THE UNIVERSITY
of EDINBURGH

Genome evolution in the genus *Caenorhabditis*

Lewis Stevens

Submitted for the degree of Doctor of Philosophy

Institute of Evolutionary Biology

University of Edinburgh

2020

Authorship declaration

I declare that this thesis is my own work, and that the work described here is my own except where explicitly stated. This work has not been submitted for any other degree or professional qualification.

Lewis Stevens

January 2020

Acknowledgements

Firstly, I thank Mick Watson, my second supervisor, and Ally Phillimore, my committee member, for their support and guidance over these last few years. I also thank Baillie Gifford for funding an obscure PhD student to work on a bunch of obscure worms.

None of the work presented in this thesis would have been possible without all those who have patiently sifted through foosty fruits and flowers hunting for microscopic worms. I thank all members of the *Caenorhabditis* Genomes Project, and in particular Marie-Anne Félix and Aurélien Richaud, for being so helpful, open, and collaborative. I hope the genomes prove useful.

A huge thanks must also go to my friends and colleagues in Kenya, who were only too willing to stick their hands in countless ears on the hunt for a worm they'd never heard of. And to Stefan, Ellie and the cats for letting me into their home. I hope to visit Busia again one day.

My happiness as a PhD student was, in large part, due the relationships I had with past and present members of the Blaxter lab (Georgios, Sujai, Dom, Laura, Elisabet, Carlos, Reuben, Andrea, Pablo, and many others). It's rare to find such a fantastic group of people in a single lab group and I'm immensely thankful for their guidance and patience, particularly in those early days. Special thanks must go to Richard Challis who had the thankless task of keeping the Blaxter lab computer cluster up and running. You may feel your efforts often go unnoticed, but we are all incredibly grateful.

Thanks to the people of Ashworth for creating such a fun and welcoming research environment. Many of the finest people I know spend their days sipping coffee and chatting nonsense in the DDH. Thanks, in particular, to the TYD crew for showing me that being a good scientist doesn't mean that you have to have your life together. Special thanks to my flatmates, Mary and Koo, for ensuring that my last year in Edinburgh was a happy one. I promise I'll do the dishes once this thing is printed. A special mention must also go to Charlotte for her patience over last few months – I will make it up to you.

Thanks also to my parents for their never-ending love and support (both financial and otherwise). Given that it seems I'll be staying in academia, it's unlikely I'll ever be able to pay you back. Sorry.

Lastly, I thank my supervisor, Professor Mark Blaxter. For reasons that remain obscure, Mark decided to give me an opportunity in his lab when I was a clueless teenager destined for a career outside of Academia. I doubt there are many other supervisors that would've shown the patience and enthusiasm that Mark has over these last few years. His commitment to open and collaborative science, and, more importantly, to being a decent human being, is something I will aspire to throughout my career. I would not be doing science if it wasn't for the opportunities you have given me and I am eternally grateful.

Abstract

The free-living nematode *Caenorhabditis elegans* is an important laboratory model organism that has been fundamental to the understanding of metazoan biology. However, until recently, *C. elegans* has largely been studied in isolation, with its evolutionary history poorly understood. Recent progress in our understanding of natural ecology of *C. elegans* has led to the discovery of many new species from the genus *Caenorhabditis*, most of which are in laboratory culture. In 2014, an international collaboration was launched which aimed to sequence the genomes of all species currently in culture. In this thesis, I present draft genomes for 38 *Caenorhabditis* species generated using both short- and long-read technology. I also present the genome of *C. bovis*, a species which appears to live parasitically in the ears of cattle in Eastern Africa. I exploited these and other genome sequences to perform the most comprehensive reconstruction of the *Caenorhabditis* phylogeny to date. Analysing genome size and content within the context of this phylogeny, I reveal extensive variation in genome size that is driven by changes in gene number and repetitive content. The work presented in this thesis represents a substantial contribution to the understanding of genome evolution in *Caenorhabditis*. Moreover, these data will become fundamental in our attempts to understand the evolutionary origins of this important model organism.

Lay summary

Caenorhabditis elegans is a tiny, free-living nematode, or roundworm, which is used extensively in biological research. An improved our understanding of the natural habitat of *C. elegans* in recent years has led to the discovery of many closely related species from across the world. By studying these new species alongside *C. elegans*, it is hoped that we will understand how the biology of this important model organism evolved. In this thesis, I present genome sequences for many of these newly discovered species. I also present the genome of one species that, unlike the others, appears to live a parasitic lifestyle in the ears of cattle in Eastern Africa. I show how these genomes can be exploited to understand how *C. elegans* relates to these new species. I also investigate how the size and content of these genomes has changed over evolutionary time. The resources presented in this thesis will be essential in future attempts to understand the evolutionary origins of this important nematode and in the study of the evolutionary forces that drive genomic change.

Table of contents

List of tables & figures

Chapter 1

General Introduction

Thesis overview

In this thesis, I present the results of a genus-wide genome sequencing project of the genus *Caenorhabditis*. I present draft genomes for 38 *Caenorhabditis* species generated using best-practice approaches. I use these genomes, along with those produced by other laboratories, to infer the phylogenetic relationships in the genus. I investigate patterns of genome evolution in the context of this phylogeny. This work represents a significant contribution to our understanding of genome evolution in this important genus of nematodes. Moreover, the data presented in this thesis will become an important resource for understanding the evolutionary origins of *Caenorhabditis elegans,* an important model organism. In this introduction, I provide the background and justification for this work.

Caenorhabditis elegans and the need for an evolutionary context

Caenorhabditis elegans as a model organism

The free-living, bacterivorous nematode *Caenorhabditis elegans (Maupas 1900)*, colloquially known as 'the worm', is a key laboratory model organism that has been fundamental to the understanding of metazoan biology. In the early 1960s, Sydney Brenner proposed the use of *C. elegans* as a laboratory model for animal development and neurobiology (Riddle *et al.* 2011). *C. elegans* is an attractive experimental model organism for several reasons: adult *C. elegans* are small (~1 mm), have a short-generation time (3 days), are transparent, and can easily be grown in large numbers on petri dishes seeded with *Escherichia coli* (Riddle et al. 2011). *C. elegans* primarily reproduces *via* self-fertile hermaphrodites, but hermaphrodites can be cross-fertilised by rare males permitting the use of genetic crosses (Brenner 1974). Despite its simplicity, *C. elegans* shares much of its biology with other animals, including humans (Corsi *et al.* 2018). As a result, *C. elegans* has become one of the most widely used experimental model systems in modern biology. Several key discoveries have been made using *C. elegans*, including the discovery of RNA interference (RNAi) (Fire *et al.* 1998), and a vast body of knowledge has accumulated about its biology, including the complete embryonic cell lineage (Sulston *et al.* 1983) and a complete map of its neuronal circuitry (the "connectome") (White *et al.* 1986; Cook *et al.* 2019b). In 1998, *C. elegans* became the first metazoan to have its genome completely sequenced (*C. elegans* Sequencing Consortium 1998).

The *C. elegans* genome

The 100 Mb *C. elegans* genome consists of five autosomes (I-V) and single sex chromosome (the X) that encode ~20,000 protein-coding genes (Brenner 1974; *C. elegans* Sequencing Consortium 1998). Several gene families are notably expanded in

the *C. elegans* genome relative to other metazoans, including seven-transmembrane G-protein coupled receptors (7TM-GCPRs) which constitute ~7% of the *C. elegans* gene set (Robertson 1998). Repetitive sequences, including transposable elements, constitute ~17% of the C. elegans genome (Sulston & Brenner 1974). Unusually for a eukaryote, a substantial proportion (15%) of genes in the *C. elegans* genome are organised in operons, where two or more genes share the same promoter (Spieth *et al.* 1993). These genes are transcribed as a single polycistronic mRNA which is subsequently resolved *via* trans-splicing (Spieth *et al.* 1993). *C. elegans* chromosomes are holocentric and therefore lack defined centromeres (Albertson & Thomson 1982). The five autosomes show a non-random distribution of genomic features, with distinct domains termed 'arms' and 'centers' which differ in their composition (*C. elegans* Sequencing Consortium 1998). The centers are enriched for deeply conserved genes and operons and have relatively few transposable elements, while the arms have fewer operons and a greater density of repeats (*C. elegans* Sequencing Consortium 1998). This structure is highly correlated with recombination rate, with the centers having substantially lower rates of recombination than the arms (Rockman & Kruglyak 2009). Relative to other organisms for which recombination rates have been studied, *C. elegans* appears to be unusual in having broad regions or domains with fairly constant recombination rate. In humans, mice, *Drosophila*, and *Arabidopsis*, recombination tends to be restricted to small regions of the genome termed recombination "hotspots" (Myers et al. 2010; Baudat et al. 2010; Comeron et al. 2012; Yelina et al. 2012).

A new evolutionary context for *C. elegans*

Despite years of scrutiny in the laboratory, details of the ecology of *C. elegans* outside the laboratory were extremely scare for many years (Félix & Braendle 2010). The N2 strain used in laboratories across the world was isolated from a compost heap in Bristol, UK, and until recently *C. elegans* was commonly thought of as a "soil nematode", despite being rarely recovered from soil samples (Félix & Braendle 2010). In 2012, studies of natural populations revealed that *C. elegans* naturally thrives in far more microbe-rich environments such as rotting fruits (Félix & Duveau 2012).

This new ecological understanding, combined with extensive worldwide sampling efforts, led to the isolation of numerous wild strains of *C. elegans* (Andersen *et al.* 2012; Cook *et al.* 2017; Ferrari *et al.* 2017; Richaud *et al.* 2018). In concert, in the last decade, over 50 new species from the genus *Caenorhabditis* have been discovered, all of which are in laboratory culture (Kiontke *et al.* 2011; Félix *et al.* 2014; Ferrari *et al.* 2017; Stevens *et al.* 2019).

Caenorhabditis species are now known to be found in decaying fruits, flowers and plant material worldwide (Kiontke *et al.* 2011; Félix *et al.* 2014; Ferrari *et al.* 2017; Stevens *et al.* 2019). Closely related species are often morphologically indistinguishable, and this has motivated the use of DNA sequence-based approaches for species delineation (Kiontke *et al.* 2011; Félix *et al.* 2014). Many species, including *C. elegans*, are known to have phoretic associations with invertebrate transport hosts (Kiontke & Sudhaus 2006). For some, such as *C. drosophilae* which is transported between rotting saguaro cacti on the heads of *Drosophila nigrospiracula (Kiontke 1997)*, this association appears to be highly specific. For others, such as *C. elegans,* which has been isolated from snails, millipedes and dipterans, several different transport host species are used (Kiontke & Sudhaus 2006). Other than *C. elegans,* only two other species in the genus reproduce *via* self-fertilisation, with self-fertile hermaphroditism having independently evolved from obligately outcrossing gonochoristic ancestors three times (Kiontke *et al.* 2004, 2011). Unlike *C. elegans*, populations of outcrossing *Caenorhabditis* species have extremely high levels of nucleotide diversity, suggestive of large effective population sizes (Cutter *et al.* 2006; Dey *et al.* 2013). Indeed, wild *C. brenneri* populations were found to contain the highest levels of genetic diversity of any known eukaryote (Dey et al. 2013). Surprisingly, despite limited morphological divergence, the genetic distance spanned by the genus is large, with the closely related species *C. elegans* and *C. briggsae* being more genetically distinct than humans and mice (Kiontke *et al.* 2004). In 2003, the genome of *C. briggsae* was sequenced, enabling the first comparative genomics studies in *Caenorhabditis (Stein et al. 2003).* Comparisons between the genomes of *C. briggsae* and *C. elegans* revealed that synteny is poorly conserved (Stein et al. 2003;

Hillier et al. 2007), with a rate of intrachromsomal rearrangement that is fourfold higher than reported for *Drosophila* (Coghlan and Wolfe 2002). Despite this, genes have largely remained on the same chromosome over large spans of evolutionary time, with little interchromosomal rearrangement (Stein *et al.* 2003; Hillier *et al.* 2007). Since then, the genomes of several other *Caenorhabditis* species have been published and have been used to reconstruct the *Caenorhabditis* phylogeny, facilitate species description, and investigate the genomic consequences of reproductive mode (Mortazavi *et al.* 2010; Fierst *et al.* 2015; Slos *et al.* 2017; Kanzaki *et al.* 2018; Yin *et al.* 2018; Stevens *et al.* 2019).

The *Caenorhabditis* Genomes Project

In 2014, seeking to exploit this newly discovered *Caenorhabditis* diversity, several members of the *Caenorhabditis* research community initiated the *Caenorhabditis* Genomes Project (CGP), an international collaboration which aimed the sequence the genomes of all species currently in culture (caenorhabditis.org). These data, it was hoped, would help to provide an essential evolutionary context to *C. elegans* and the vast body of associated research. Realising the aims of this project has formed the basis of the work presented in this thesis.

Sequencing genomes

Historical overview

Following the development of the first RNA sequencing methods in the late 1960s (Brownlee *et al.* 1967; Barrell & Sanger 1969), the first method of determining the sequence of bases in a DNA molecule was published in 1977 (Sanger *et al.* 1977b). This approach, known as Sanger sequencing, uses chain-terminating dideoxynucleotides and led to the publication of the first complete genome sequence, that of the bacteriophage φX174 (Sanger *et al.* 1977a). In 1998, the first genome of a multicellular organism, *C. elegans* (*C. elegans* Sequencing Consortium 1998*)*, was sequenced, laying the foundations for future genome sequencing projects, including the Human Genome Project (Lander *et al.* 2001). In the early 2000s, next-generation sequencing (NGS) technologies became available that were capable of producing huge quantities of data relatively cheaply (Goodwin *et al.* 2016). More recently, new sequencing technologies, so called 'third generation' technologies, have become available that offer substantially longer read lengths (Sedlazeck *et al.* 2018). Current DNA sequencing technologies are typically categorised by whether they produce short reads (< 1 kbp) or long reads (>1 kbp).

Short-read sequencing

Prior to 2010, several short-read sequencing technologies were available, including Illumina Solexa, Ion Torrent and Roche 454 sequencing. However, Illumina's 'sequencing by-synthesis' technology now dominates the short-read sequencing market (Mohamed & Syed 2013). Briefly, this approach involves detecting the incorporation of fluorescently-labelled nucleotides (dNTPs) into DNA fragments bound to a flow cell (Goodwin *et al.* 2016). This method of DNA sequencing is highly accurate, with current platforms having an error rate of less than 1 in 10,000 (99.99% accuracy) (Schirmer *et al.* 2015). The resulting reads are between 50 and 300 bp in length depending on platform. Both ends of a DNA fragment can be sequenced during the same run, resulting in 'paired-end' reads. The highly parallel nature of this

technology means that Illumina platforms are capable of producing huge quantities of data at relatively low cost (Goodwin *et al.* 2016). Wide adoption of this technology has led to massive increase in the number of draft genome sequences available (Goodwin *et al.* 2016).

The large number of reads generated from short-read sequencing platforms prevent the use of assembly strategies that involve finding overlaps between reads directly. Instead, reads are typically assembled using a de Bruijn graph approach (Compeau *et al.* 2011). Briefly, this involves extracting short sequences from the reads, known as kmers, and storing their sequences and frequencies. Overlaps between kmers are identified and used to create a de Bruijn graph. This graph is then traversed and any unambiguous paths through the graph are used to create contigs. If paired-end information is available, contigs can be joined together to form scaffolds. Many implementations of this approach exist (e.g. SPAdes (Bankevich *et al.* 2012), Velvet (Zerbino & Birney 2008) and ABySS (Simpson *et al.* 2009)), including some specifically designed for the assembly of heterozygous genomes (eg. Platanus (Kajitani *et al.* 2014)). However, a major drawback of this approach and of short-read sequencing generally is that sequences that are present in more than one copy in a genome (repeats), and that are longer than the read (or insert) length, cannot be resolved unambiguously. As a result, assemblies generated using short-read data are often highly fragmented, particularly those of large, repeat-rich genomes (Thomma *et al.* 2016).

Long-read sequencing

Recently, single-molecule, or third-generation, technologies offered by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) have become available. Briefly, PacBio's Single-Molecule Real-Time (SMRT) technology works by detecting the incorporation of fluorescently-labelled nucleotides into a single molecule of DNA being copied from a single-stranded template (Roberts *et al.* 2013). ONT's nanopore-based sequencing technology works by detecting disruptions in electrical current as a single DNA strand passes through a protein nanopore

embedded in a membrane (Jain *et al.* 2016). The electrical disruption is sequence-dependent and can therefore be translated into nucleotide sequence using hidden Markov models (HMMs) or recurrent neural networks (RNNs) (Wick *et al.* 2019). Both technologies are capable of producing read lengths of many kilobases in length (Jain *et al.* 2018). However, in contrast to short-read sequencing, long-read sequencing technologies have high error rate (5-15%) (Watson & Warr 2019).

Unlike short reads, long reads can be assembled by finding overlaps in the reads themselves. Overlapping reads are used to form contigs, whose sequence is composed of a consensus of all overlapping reads. Several long-read assemblers are available, including Canu (Koren *et al.* 2017), flye (Kolmogorov *et al.* 2018), and wtdbg2 (Ruan & Li 2019). However, due to the high error rate of long reads, assemblies generated using long reads alone typically contain large numbers of sequencing errors which must be corrected prior to downstream annotation and analysis (Watson & Warr 2019). To address this, highly-accurate short-reads can be aligned to the long-read assembly to correct remaining sequencing errors (Walker *et al.* 2014). As longer reads can span longer repetitive sequences, assemblies generated using long-reads are typically substantially more contiguous than those generated using short reads (Roberts *et al.* 2013).

High-throughput mapping approaches

While assemblies generated using long-reads are typically highly-contiguous, long-reads alone are often insufficient to assemble contigs or scaffolds that represent complete chromosomes. Previously, ordering and orienting contigs or scaffolds into chromosomes was achieved using genetic mapping (Davey *et al.* 2011). While this approach is relatively straightforward and inexpensive, performing genetic crosses is not feasible for many organisms. Recently, technologies (eg. BioNano, 10X and HiC) capable of providing very long range information without the need for genetic mapping have been developed. In conjunction with long-reads, these approaches have been used to produce chromosome-scale draft genomes (Cotton *et al.* 2016).

Genomics of non-model organisms

Generating high-quality draft genomes of non-model organisms remains challenging. Unlike model organisms, which are often highly-inbred, the genomes of many non-model organisms are heterozygous which can complicate *de novo* assembly (Vinson *et al.* 2005; Barrière *et al.* 2009). In organisms which are easily inbred, such as *Caenorhabditis* species, this issue can be circumvented by generating inbred lines prior to sequencing. However, inbreeding is not feasible for many organisms. In addition, as much of life on earth is small, the amount of DNA that can be extracted from a single individual is often insufficient for DNA sequencing. Therefore, it is necessary to extract DNA from multiple individuals in a population, and this will complicate *de novo* assembly if the population is genetically polymorphic (Nowell *et al.* 2018). Protocols capable of amplifying small amounts of DNA recovered from a single individual, known as whole-genome amplification, have been developed, but these approaches generate chimeric DNA fragments, leading to misassemblies (Lasken & Stockwell 2007; Sabina & Leamon 2015). Recently, protocols for generating long-read sequencing datasets using DNA from single individuals have been developed (Kingan *et al.* 2019), promising to allow long-read sequencing to be used for many other organisms from across the tree of life. In addition, small organisms are also hard to separate from non-target organisms, such as bacteria and fungi. As a result, sequencing datasets of these organisms often contain significant amounts of contaminating reads which must be removed prior to downstream analysis. Several approaches for identifying and removing contaminant reads have been developed, including blobtools, which uses taxon-annotated GC-coverage plots to identify reads from non-target organisms (Laetsch & Blaxter 2017a). In 2017, the genome of the tardigrade *Hysibius exemplaris* was published alongside claims that a significant fraction of its genome had been horizontally transferred from various bacterial and fungal species (Boothby *et al.* 2015). This claim was almost immediately revealed to be an artefact of assembly contamination (Arakawa 2016; Bemm *et al.* 2016; Koutsovoulos *et al.* 2016).

Chapter 2

De novo assembly of the genomes of 38 *Caenorhabditis* species

A proportion of the work presented in this chapter was carried out by people other than myself. There contributions are listed in Table 1. As a result, first person plural is used throughout.

Preface

In this chapter, we present draft genome sequences for 38 *Caenorhabditis* species generated as part of the *Caenorhabditis* Genomes Project. Using a range of numerical and biological metrics, we assess the quality of these data along with data for the genomes of a further 18 species generated in other laboratories and discuss their suitability for downstream analyses.

Results

Draft genome sequences for 38 *Caenorhabditis* species

We sequenced the genomes of 40 *Caenorhabditis* species to high coverage (150-400x), from inbred strains where possible (Table S1), using short-read Illumina platforms. We also sequenced the genomes of five species (*C. bovis*, *C. guadeloupensis*, *C. monodelphis*, *C. portoensis*, and *C. vivipara*) to 30-350x coverage using the Oxford Nanopore MinION or PromethION long-read platforms. We assembled each genome *de novo* using either short-read or long-read specific pipelines (Fig. S1A, B). Reads originating from non-target organisms were identified in preliminary assemblies and discarded. The genome assemblies for *C.* sp. 45 and *C.* sp. 47 were not of sufficient quality for protein-coding gene prediction (N50 lengths < 6kb), likely due to high levels of heterozygosity, and are not discussed further[1].

The contiguity of the resulting assemblies varies (Table 1; Fig. 1A,B), with long-read assemblies being considerably more contiguous than short-read assemblies (mean N50 of 3.3 Mb and 96.8 kb, respectively; Fig. 1C; Fig. S2A). N50 lengths of the short-read assemblies range from 15.1 kb (*C. waitukubuli*) to 224.5 kb (*C. tribulationis*), comprising 21,203 and 3,278 scaffolds, respectively. Of the five long-read assemblies, the assembly of *C. bovis* is the most contiguous, comprising just 35 contigs with an N50 of 7.5 Mb. It is likely that heterozygosity impacted the contiguity of both the short and long read assemblies (Fig. S2B). The two least contiguous short-read assemblies (*C. waitukubuli* and *C.* sp. 46) and the two least contiguous long-read assemblies (*C. guadeloupensis* and *C. vivipara*) were all of non-inbred strains (Table S1). The high contiguity of the *C. bovis* assembly, despite that fact it was not deliberately inbred, suggests that this isolate may have had naturally low heterozygosity (discussed in chapter 3). Despite these differences in contiguity, all

[1] For both species, we predicted protein sequences from transcriptome assemblies to permit their inclusion in downstream analyses.

38 assemblies are highly complete, with a mean of 95.6% complete BUSCO genes present in each (Fig. 1C).

Assembly span is also highly variable, with greater than five-fold variation across the 38 assemblies (Table 1; Fig. 1A; Fig. 2A,B). The smallest assembly (*C. drosophilae*) spans 48.5 Mb, while the largest (*C. vivipara*) spans 249.2 Mb (Table 1). However, assemblies of heterozygous genomes can contain regions of uncollapsed heterozygosity, which result in assembly spans that are substantially larger than the haploid genome size (Barrière *et al.* 2009). We used the proportion of duplicated BUSCO genes to estimate the amount of uncollapsed heterozygosity in each assembly. The majority of assemblies contain low numbers of duplicated BUSCO genes (0.5 - 2.4%; Table 1; Fig. 2A,B) and are comparable to *C. elegans* (0.6%), suggesting that assembly span is an accurate estimate of haploid genome size in these species. It is important to note, however, that, as BUSCO genes are highly conserved, they are likely to have lower levels of polymorphism which will lead us to underestimate the degree of duplication in our assemblies. Assembles for five species (*C. guadeloupensis, C.* sp. 46, *C.* sp. 49, *C. waitukubuli*, and *C. vivipara*) contain over 4% of BUSCO genes in multiple copies (Fig. 2B). The spans of these assemblies are therefore likely to be larger than the respective haploid genome span.

Species	*Span (Mb)*	*Number of contigs*	*N50 (kb)*	*Longest contig (Mb)*	*N%*	*GC content (%)*	*Read type*	*Protein-coding genes*	*BUSCO complete % (genome/gene set)*	*BUSCO duplicated % (genome/gene set)*	*BUSCO fragmented % (genome/gene set)*
C. afra	65.62	1898	176.07	1.05	0.31	47.1	short	19,834	97.3/96.8	0.9/1	2.3/2.5
C. astrocarya	50.5	3479	146.2	0.97	0.17	34.7	short	14,030	95.1/95.2	0.5/0.7	4/3.8
C. bovis	62.73	35	7,557.46	10.86	0	38.1	long	13,128*	94.2/95.1	1.6/1.7	4.6/3.2
C. castelli	77.88	1964	161.5	0.91	1.05	33.9	short	19,694	93.8/94.4	0.8/1.3	4.6/4.2
C. dolens	110.77	6411	86.48	0.57	0.65	35.4	short	26,684	95.1/93.9	2.2/2.5	3.9/4.6
C. doughertyi	140.72	10147	69.88	0.5	0.11	37.4	short	30,860	97.5/98.4	2.1/2.1	1.9/1.2
C. drosophilae	48.49	2404	91.72	0.49	0.12	33.1	short	13,712	96.2/94.8	1.1/1.4	3.4/4.3
C. guadeloupensis	121.34	1214	439.55	3.71	0	34.3	long	23,983	95.7/95.5	8.7/8.7	2.4/2.9
C. imperialis	69.38	3672	53.58	0.46	0	42.4	short	18,588*	96/93.2	0.8/0.9	3.1/4.4
C. macrosperma	92.52	6369	70.37	0.85	0.07	44.8	short	24,184	97.7/97.3	1/1.3	1.4/2
C. monodelphis	117.11	125	2,970.97	8.28	0	44.1	long	17,803	91.7/93.2	1.5/2.2	6/5.2
C. nouraguensis	72.89	1915	153.11	0.77	0.61	43.0	short	22,774	97/97	0.9/1.4	2/2.1
C. oiwi	91.71	4417	116.06	0.94	0.14	37.5	short	23,208	96.9/91.6	2.3/2.4	1.8/5.7
C. parvicauda	93.74	5719	44.41	0.31	1.11	39.3	short	16,412	89.7/89.9	1.4/2.3	5.7/6.4
C. plicata	111.46	5064	58.29	0.3	0.15	33.3	short	14,843	90.6/93.7	0.8/0.7	5.6/4.3
C. portoensis	161.47	541	5,023.04	27.69	0	35.2	long	22,506	96.2/96.5	1.6/1.5	2.6/2.6
C. quiockensis	100.37	4890	139.43	0.93	0.09	34.8	short	22,238	96.1/95.6	1.6/2.3	3.2/3.4
C. sp. 2	49.93	2773	78.13	0.41	0.13	31.1	short	13,557	96.3/96.7	0.9/0.8	3.2/2.9
C. sp. 8	66.77	2319	88.47	0.57	0.1	36.5	short	16,224	95.1/94.5	1.1/1.2	3.9/4.6
C. sp. 24	74.48	3997	84.9	0.48	0.75	36.3	short	18,430	94.7/94.6	1.3/2	4.2/4.1
C. sp. 25	74.12	4063	52.95	0.38	0.13	40.2	short	20,027	96/95.9	0.7/0.7	2.7/2.9
C. sp. 27	160.76	7032	100.21	0.67	0.99	37.1	short	20,244	94.7/95.1	0.6/1	3.9/3.4
C. sp. 30	57.07	3340	57.23	0.37	0.15	37.1	short	15,211	95.7/95.3	1/1.2	3.5/3.7
C. sp. 46	99.88	13383	27.06	0.59	0.65	43.2	short	29,209	95.3/94.3	7.5/8.4	3.4/4.2
C. sp. 48	132.61	4122	153.22	2.1	0.12	38.0	short	32,633	97.9/97.4	2.4/2.6	1.5/1.7
C. sp. 49	77.94	3998	110.56	0.6	0.12	42.2	short	24,004	96.8/95.8	5.8/5.8	2.4/2.9
C. sp. 51	91.12	2703	104.21	0.52	0.13	37.7	short	23,345	97.9/96.5	1.2/0.9	1.4/2.2
C. sp. 54	165.38	12665	49.56	0.54	0.33	35.4	short	35,881	97.6/96.6	1.9/1.8	1.8/2.7
C. sp. 55	112.82	5908	92.43	0.76	0.05	38.3	short	26,129	97.6/97.8	1.5/2.3	1.9/1.7
C. sp. 56	91.49	3767	75.19	0.62	0.13	36.6	short	19,698	94.6/94.2	1.4/1.8	4.3/4.3
C. sulstoni	65.1	2044	136.67	1.01	0.59	46.4	short	18,192	98/95.3	0.5/0.9	1/3.3
C. tribulationis	101.23	3276	224.52	1.2	0.15	41.5	short	24,787	97.7/97.9	1.5/1.7	1.1/1.4
C. uteleia	104.05	3222	177.02	1.91	0.99	36.9	short	27,614	96/96.1	1.9/2.1	3.1/3.1
C. virilis	93.43	7132	35.85	0.22	0.01	33.5	short	19,447	92.5/92.6	1.9/2.2	5.2/5.4
C. vivipara	249.22	3253	500.6	8.78	0	36.3	long	25,568	93/93.4	6.8/7.7	4.5/4.8
C. waitukubuli	91.42	21203	15.12	0.84	1.27	41.9	short	30,089	92.6/92.8	4.3/5.2	4.7/5.5
C. wallacei	80.67	3304	71.82	0.47	0.13	36.4	short	20,860	97.9/97.2	0.5/0.7	1.7/2
C. zanzibari	101.09	3128	91.33	0.51	0.22	38.3	short	22,198	98/98.5	1.3/1.9	1.4/1
C. elegans	100.2	7	17,493.8	20.92	0	35.4	-	20,208	98.7/98.7	0.6/0.4	0.7/0.9

Table 1: Genome and gene set metrics for *de novo* genome assemblies from 38 *Caenorhabditis* species.

Metrics for the *C. elegans* N2 reference genome are shown for comparison. Assembly and gene set completeness were assessed using BUSCO (version 3.0.2) with the 'nematoda_odb9' dataset. *RNA-seq data were not available and were therefore not used during prediction.

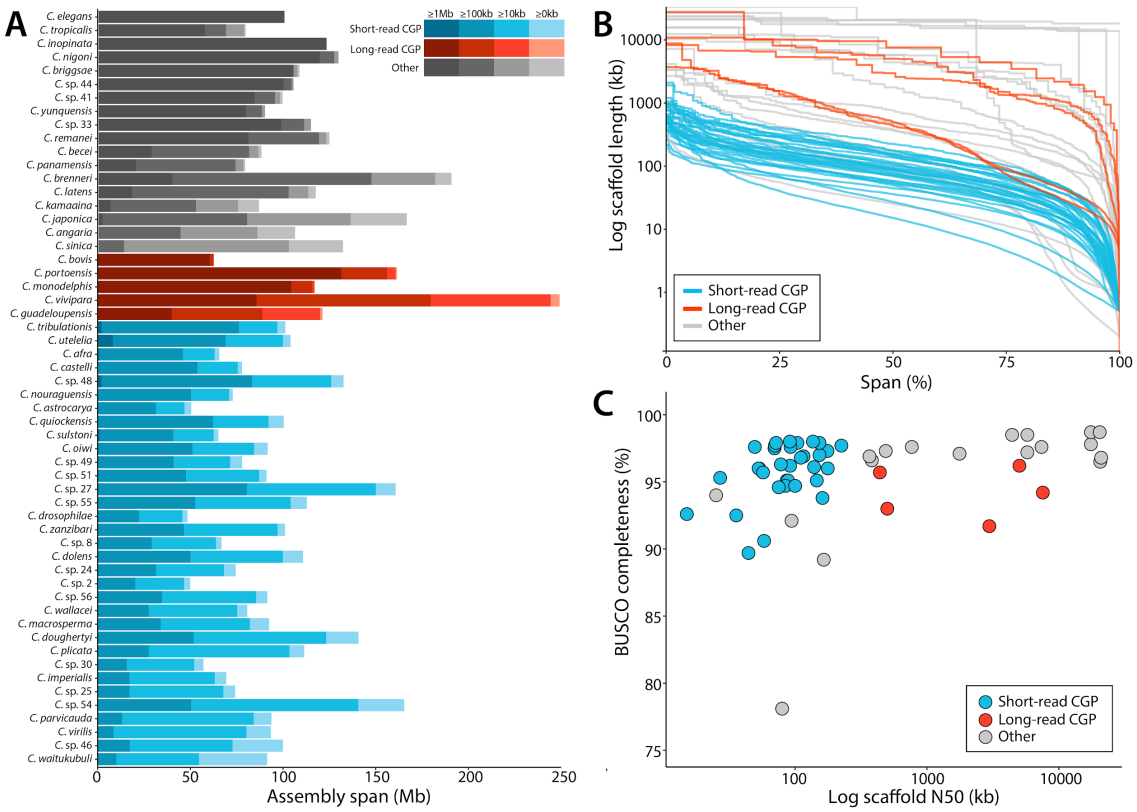

Figure 1: Contiguity and completeness of 56 *Caenorhabditis* draft genomes.

A: Proportion of each assembly in scaffolds of different lengths. Each group (short-read, long-read and other) is ordered by N50 length. **B:** Cumulative scaffold lengths as a proportion of assembly span. **C:** Genome BUSCO completeness compared with caffold N50 length.

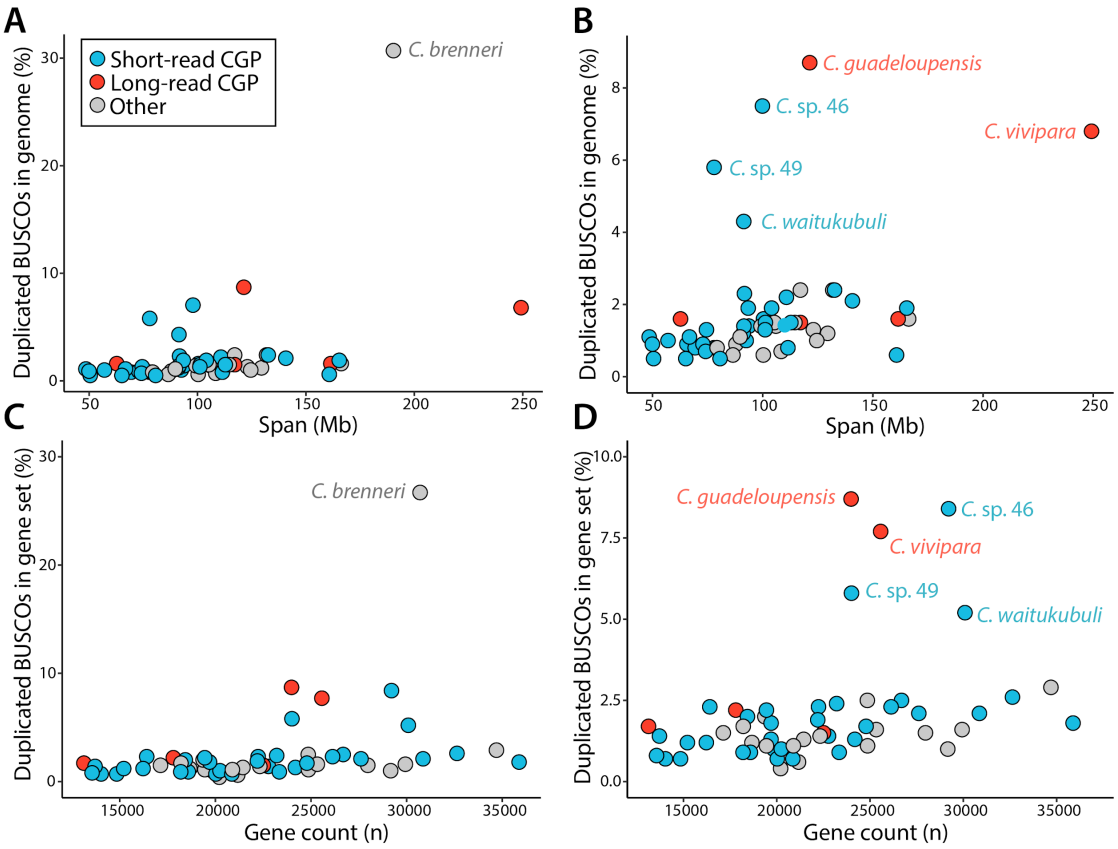

Figure 2: Duplication in 56 *Caenorhabditis* draft genomes and gene sets.

A: BUSCO gene duplication assessed in the genomes of all species. **B:** BUSCO gene duplication assessed in the genomes of all species except *C. brenneri* (30.7% duplicates). Species names of assemblies with high levels of duplication are shown. **C:** BUSCO gene duplication assessed in the gene sets in all species. **D:** BUSCO gene duplication assessed in the gene sets in all species except *C. brenneri* (26.7% duplication). Species names of assemblies with high levels of duplication are indicated.

Protein-coding gene sets of 38 Caenorhabditis species

To assist with protein-coding gene prediction, we sequenced the transcriptomes of all species (except *C. bovis* and *C. imperialis*) to high coverage using short-read Illumina platforms. We aligned the RNA-seq reads to each genome assembly and provided the resulting alignments to BRAKER to generate gene predictions (Fig. S1C). For *C. bovis* and *C. imperialis*, we aligned protein sequences from six related nematode species to both genome assemblies and provided the inferred intron positions to BRAKER to guide gene predictions (Fig. S1C).

The number of protein-coding genes predicted in each genome varies considerably (Table 1), with over three-fold variation across the 38 gene sets. The smallest gene set (*C. bovis*) contained 13,128 genes, while the largest (*C.* sp. 54) contained 35,881 genes. To assess the completeness of each predicted gene set, we compared the proportion of BUSCO genes found in each gene set with the proportion found in the respective genome assembly. Gene set completeness was highly correlated with genome completeness (r^2=0.6245, $P < 0.001$; Fig. 3A), with the majority of gene sets containing 0-1.5% fewer BUSCO genes than their respective assemblies. *C. imperialis, C. oiwi*, and *C. parvicauda* were notable outliers, however, with each gene set containing >2% fewer BUSCO genes than the respective assemblies (Fig. 3A). We also assessed the proportion of fragmented genes in each gene set, with a high number of fragmented genes relative to the genome assembly potentially indicating a high proportion of fragmented gene models. The majority of gene sets contain <1.5% more fragmented genes than their respective genomes (Fig. 3B). *C. oiwi* and *C. imperialis* were notable outliers, with 3.9% and 2.3% more fragmented genes in their gene sets than the respective genome assemblies (Fig. 3B).

As for the genome assemblies, we used the proportion of duplicated BUSCO genes as an estimate of overall duplication in each gene set. The majority of gene sets, like their respective genomes, had low levels of duplication (0-2.6%), suggesting that variation in number of predicted genes represents biological variation in gene number in the genus rather than a technical artifact (Fig. 2C,D; discussed in Chapter

5). As expected, the gene sets of all five species identified previously as containing high levels of uncollapsed heterozygosity contain an excess of duplicated genes (5.2-8.7%; Fig. 2C,D). The number of genes predicted for these species is therefore is likely to be an overestimate of the number of genes present in their genomes.

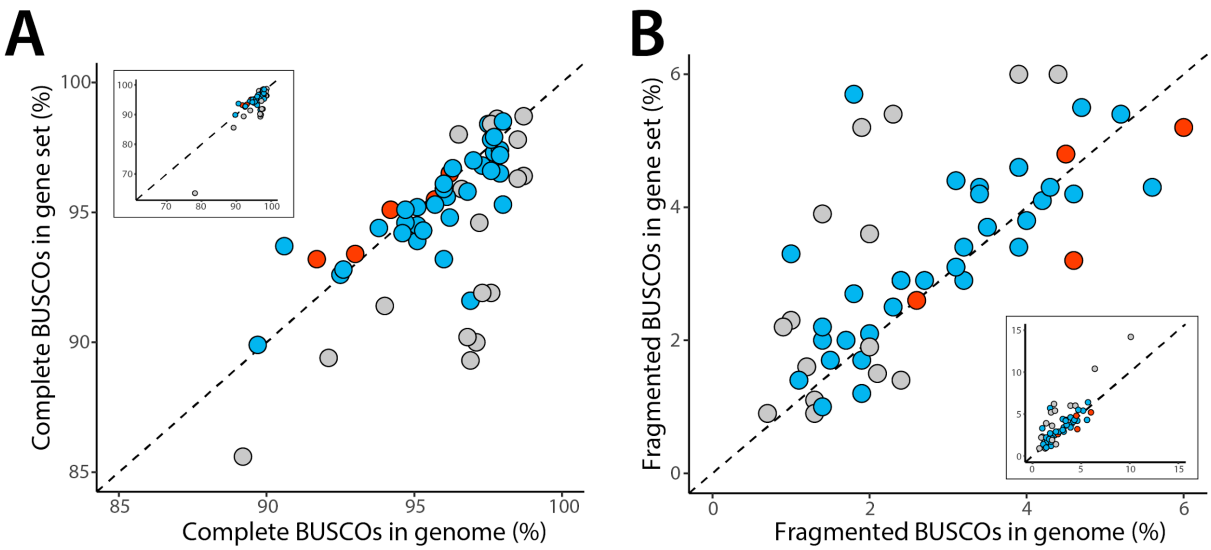

Figure 3: Relative completeness and fragmentation of BUSCO reference genes in the gene sets of 56 *Caenorhabditis* species compared to their genomes.

A: Completeness of BUSCO genes in genome assemblies and gene sets in all species except *C. angaria*. Inset: all species. **B:** Fragmentation of BUSCO genes in genome assemblies and gene sets of all species except *C. angaria* and *C. japonica*. Inset: all species.

Discussion

In this chapter, I have presented draft genome sequences and protein-coding gene sets for 38 *Caenorhabditis* species. Using both numerical and biological metrics, I demonstrate that the majority of these data are of high quality, highly-complete, and suitable for downstream analyses. I found that assemblies generated from long-reads were significantly more contiguous than those generated from short-reads. After accounting for assembly and annotation artefacts, I also found substantial variation in assembly span and predicted gene count that appear to be the result of real biological variation within the genus (discussed in Chapter 5). These data, combined with genomes generated in other laboratories, represent draft genomes and gene sets for 56 of the 64 *Caenorhabditis* species currently in culture (Kiontke *et al.* 2011; Félix *et al.* 2014; Ferrari *et al.* 2017; Stevens *et al.* 2019); Marie-Anne Felix, Lise Frezal, Matthew Rockman, Christian Braendle, John Wang, Michael Ailion, Erik Andersen, Asher Cutter, pers. comm.).

We attempted to avoid assembly issues arising from heterozygosity by sequencing inbred strains and, as a result, a majority of our draft genomes contain low levels of duplication. However, inbred strains were not available for six species and their resulting draft genomes contain higher-than-average levels of duplication. Heterozygosity in these species also considerably impacted assembly contiguity. The least contiguous long- and short-read assemblies are of non-inbred strains, and the assemblies of two species (*C.* sp. 45 and *C.* sp. 47) were too fragmented for gene prediction. The difficulty of generating high-quality assemblies of highly heterozygous genomes is well-known (Vinson *et al.* 2005; Takeuchi *et al.* 2012; You *et al.* 2013). This problem is particularly acute for outcrossing *Caenorhabditis* species which contain extremely high levels of nucleotide diversity (Dey *et al.* 2013). Several previously published genomes of outcrossing *Caenorhabditis* species are known to be highly duplicated, with 10-30% of genes present in multiple copies (Barrière *et al.* 2009). As inbreeding these nematodes is relatively straightforward and inexpensive, we suggest that improved versions of these genomes should be generated by

resequencing inbred strains. Until this goal is realised, these genomes and gene sets should be excluded from downstream analyses that concern genome size or gene count.

The majority of the data we present are of high-quality and considerably better than previously published assemblies. However, none of our assemblies equal the quality of the *C. elegans* reference genome, which is fully assembled into chromosomes and has a gene set which has undergone extensive manual curation (C. elegans Sequencing Consortium 1998; Lee *et al.* 2018). While it is unlikely that any of the species sequenced here will receive a similar level of scrutiny to *C. elegans*, advances in sequencing technology continue to substantially improve the quality of genome assembly and gene prediction. We have shown that long-read sequencing technology can significantly increase assembly contiguity and we hope that long-read data can be generated for all species in the genus. High-throughput mapping approaches, such as 10X or HiC, can be combined with long-read sequencing to order and orient large contigs into scaffolds representing full chromosomes (Cotton *et al.* 2016). Long-read RNA-seq, including direct RNA sequencing, are becoming commonplace (Wang *et al.* 2016; Zulkapli *et al.* 2017; Keller *et al.* 2018), and tools to exploit these new data for gene prediction have been released (Cook *et al.* 2019a), promising to dramatically increase gene prediction accuracy. Therefore, while our data represent a good, first estimates of the genomes of these species, we hope that they are considerably improved as sequencing technology continues to advance.

Methods

Details of software versions and parameters used in these analyses are available in Tables S2-3.

Nematode culture and nucleic acid extraction

Each strain was grown on 60 mm or 90 mm NGM plates enriched with agar (for 1 L: 3 g NaCl, 5 g bactopeptone, 10 g agar, 7 g agarose, 1 mL cholesterol 5 mg/mL in ethanol, 1 mL CaCl2 1 M, 1 mL MgSO4 1 M, 25 mL KPO4 1 M) seeded with *Escherichia coli* OP50. Nematodes were harvested just after starvation and washed in M9 supplemented with 0.001% Tween20 several times to remove *E. coli* and other contaminants. Nematode pellets were stored at -80°C until nucleic acid extraction.

To extract DNA, we resuspended 100 µL of nematode pellet in 600 µL of Cell Lysis Solution (Qiagen) and 5-20 µL of proteinase K (20 µg/uL) and incubated for 4-16 hr at 56°C. 5 µL of RNase A (20 µg/µL) was added and incubated at 37°C for 1 hr. We added Protein Precipitation Solution (Qiagen) and centrifuged at 15,000 rpm for 30 min. The supernatant was collected in a new tube and genomic DNA was precipitated by adding 600 µL of isopropanol. We centrifuged the precipitated DNA at 15,000 rpm for 3 min and discarded the supernatant. The resulting DNA pellets were washed twice with 70% ethanol and briefly allowed to dry before being resuspended in 50-100 µL of elution buffer or nuclease-free water. DNA concentration and quality were assessed using Qubit (Thermo Scientific) and agarose gel electrophoresis respectively.

To extract RNA, we resuspended 100 µL of nematode pellet in 500 µl of Trizol. The Trizol suspension was frozen in liquid nitrogen and then transferred to a 37°C water bath until completed thawed. We repeated this freeze/thaw process five times before vortexing for 5 minutes. We added 100 µL of chloroform to the Trizol suspension and mixed vigorously by hand for 15 seconds. After centrifugation (15 minutes at 15,000 rpm and 4°C), we transferred the aqueous (upper) phase to a new tube and

precipitated the RNA using 250 µL of isopropanol. The pellets were washed in 70% ethanol and briefly allowed to dry before being resuspended in 50-100 µL of RNase-free water. RNA concentration and quality were assessed using Qubit (Thermo Scientific) and gel electrophoresis respectively.

Illumina short-read sequencing

Short-insert (300-600 bp) genomic libraries and RNA-seq (insert size of 150-180bp) were prepared using Illumina Nextera reagents as per standard manufacturer's instructions by Edinburgh Genomics (Edinburgh, UK). The resulting libraries were sequenced on an Illumina platform (HiSeq 2500, HiSeq 4000, or NovaSeq) at Edinburgh Genomics.

Oxford Nanopore long-read sequencing

For each library, we sheared 2-5 µg of high molecular weight (HMW) genomic DNA to 10-35 kb using G-tubes (Covaris) or sonication and purified the sheared DNA using an AMPure XP (Agencourt) bead purification step. We then prepared libraries using '1D Genomic DNA by Ligation' (SQK-LSK108 or SQK-LSK109) as per manufacturer's instructions. We sequenced prepared libraries using MinION (FLOMIN-106) flow cells or PromethION (FLO-PRO002) flow cells. PromethION sequencing was performed at Edinburgh Genomics.

Short-read genome assembly

We performed quality control of all both genomic and transcriptomic sequence data using FastQC (Andrews & Others 2010) and used Skewer (Jiang *et al.* 2014) to remove low-quality bases and adapter sequence. We identified contaminants using taxon-annotated GC-coverage plots as implemented in blobtools (Laetsch & Blaxter 2017a). Briefly, preliminary assemblies were generated using SPAdes and likely taxonomic origin was determined by searching the nucleotide 'nt' or uniref reference proteomes databases with NCBI-BLAST+ (Camacho *et al.* 2009) or Diamond (Buchfink *et al.* 2015). Reads originating from identified contaminants were discarded. We estimated the optimal k-mer length for assembly using KmerGenie (Chikhi & Medvedev 2014)

and used JellyFish (Marçais & Kingsford 2011) and GenomeScope (Vurture *et al.* 2017) to assess kmer-spectra and estimate genome size. Assemblies were generated using several de Bruijn graph assemblers, including Velvet (Zerbino & Birney 2008), SPAdes (Bankevich *et al.* 2012) and Platanus (Kajitani *et al.* 2014), across several parameter values. The resulting assemblies were assessed using numerical metrics and BUSCO (Simão *et al.* 2015). The highest quality assembly was selected and, where possible, scaffolded using SCUBAT2 (available at https://github.com/GDKO/SCUBAT2) with transcripts assembled using Trinity (Haas *et al.* 2013).

Long-read assembly

We base-called the MinION FAST5 data using Albacore or Guppy (available at https://community.nanoporetech.com). As before, contaminants were identified using taxon-annotated GC-coverage plots as implemented in blobtools (Laetsch & Blaxter 2017a). Briefly, preliminary assemblies were generated using wtdbg2 (Ruan & Li 2019) and likely taxonomic origin was determined by searching nt or uniref reference proteomes with NCBI-BLAST+ or Diamond. Reads originating from identified contaminants were discarded. We assembled the remaining reads using wtdbg2, flye (Kolmogorov *et al.* 2018) and Canu (Koren *et al.* 2017). We assessed the resulting assemblies using numerical metrics as implemented in Quast (Gurevich *et al.* 2013) and selected the most contiguous assembly. Reads were aligned to the assembly using minimap2 (Li 2018). Sequencing errors were corrected using either Nanopolish (Loman *et al.* 2015) or Medaka (available at https://github.com/nanoporetech/medaka). Finally, we corrected any remaining sequencing errors using two iterations of Racon (Vaser *et al.* 2017) followed by two iterations of Pilon (Walker *et al.* 2014), aligning the Illumina short-reads to the resulting assembly using BWA-MEM at each iteration.

Protein-coding gene prediction

We identified repeats independently in each genome using RepeatModeler (Smit & Hubley 2010). To avoid masking protein-coding genes, we filtered out sequences which were not similar to known types of repeats in RepBase (Jurka *et al.* 2005). We combined each filtered repeat library with Rhabditida repeats obtained from RepBase (Jurka *et al.* 2005). This concatenated repeat library was then provided to RepeatMasker (Smit *et al.* 1996) for masking. If RNA-seq data were available, reads were aligned to the assembly using STAR (Dobin *et al.* 2013) and the resulting BAM file provided to BRAKER (Hoff *et al.* 2016), which performed final gene prediction. If RNA-seq data were not available, protein-coding genes were predicted using BRAKER (Hoff *et al.* 2016), using proteins sequences from nematode-specific EggNOG database (which comprises sequences from *C. elegans*, *C. briggsae*, *C. remanei*, *C. japonica, Pristionchus pacificus* and *Trichinella spiralis*) (Huerta-Cepas *et al.* 2016b) as homology evidence.

Chapter 3

The genome of *Caenorhabditis bovis*

The following chapter is a product of work by several people (information below) and has been accepted in its current form for publication in Current Biology. First person plural is therefore used throughout.

Lewis Stevens[1], Stefan Rooke[2], Laura C Falzon[3,4], Eunice M Machuka[5], Kelvin Momanyi[4], Maurice K Murungi[4], Samuel M Njoroge[4,6], Christian O Odinga[4], Allan Ogendo[7], Joseph Ogola[8], Eric M Fèvre[3,4], Mark Blaxter[1§]

1. Institute of Evolutionary Biology, Ashworth Laboratories, School of Biological Sciences, University of Edinburgh, Edinburgh, EH9 3JT, United Kingdom

2. Usher Institute, College of Medicine and Veterinary Medicine, The University of Edinburgh, Edinburgh EH9 3JT, United Kingdom

3. Institute of Infection and Global Health, University of Liverpool, 8 West Derby Street, Liverpool, L69 7BE, United Kingdom

4. International Livestock Research Institute, Old Naivasha Road, PO Box 30709 00100, Nairobi, Kenya

5. Biosciences, Eastern and Central Africa, International Livestock Research Institute (BecA-ILRI) Hub, Old Naivasha Road, PO Box 30709 00100, Nairobi, Kenya

6. Centre for Microbiology Research, Kenya Medical Research Institute, KNH Grounds, PO Box 54840 00200, Nairobi, Kenya

7. Veterinary Department, Busia County Government, PO Box Private Bag 50400, Busia, Kenya

8. Veterinary Department, Bungoma County Government, PO Box 2489 50200, Bungoma, Kenya

Abstract

The free-living nematode *Caenorhabditis elegans* is a key laboratory model for metazoan biology. *C. elegans* has also become a model for parasitic nematodes despite being only distantly related to most parasitic species. All of the ~65 *Caenorhabditis* species currently in culture are free-living, with most having been isolated from decaying plant or fungal matter. *Caenorhabditis bovis* is a particularly unusual species that has been isolated several times from the inflamed ears of Zebu cattle in Eastern Africa, where it is associated with the disease bovine parasitic otitis. *C. bovis* is therefore of particular interest to researchers interested in the evolution of nematode parasitism. However, as *C. bovis* is not in laboratory culture, it remains little studied. Here, by sampling livestock markets and slaughterhouses in Western Kenya, we successfully reisolated *C. bovis* from the ear of adult female Zebu. We sequenced the genome of *C. bovis* using the Oxford Nanopore MinION platform in a nearby field laboratory and used the data to generate a chromosome-scale draft genome sequence. We exploited this draft genome sequence to reconstruct the phylogenetic relationships of *C. bovis* to other *Caenorhabditis* species and reveal the changes in genome size and content that have occurred during its evolution. We also identified expansions in several gene families that have been implicated in parasitism in other nematode species. The high-quality draft genome and our analyses thereof represent a significant advancement in our understanding of this unusual *Caenorhabditis* species.

Introduction

The free-living nematode *Caenorhabditis elegans* is used extensively as a model for animal development, genetics and neurobiology. As the most well-studied species within the phylum Nematoda, *C. elegans* has also become a model for this extremely abundant and diverse group of animals, many of which are parasites (Blaxter *et al.* 1998; Blaxter & Koutsovoulos 2015). Attempts to understand the evolutionary origins and genetic basis of nematode parasitism often involve comparisons between parasitic nematode species and *C. elegans* (Bürglin et al. 1998; Gilleard 2004). However, *C. elegans* is only distantly related to most parasitic species which limits the efficacy of comparative studies (Blaxter & Koutsovoulos 2015). Recent years have seen significant progress in our understanding of *Caenorhabditis* diversity, with over 30 new species discovered in the last decade (Kiontke *et al.* 2011; Félix *et al.* 2014; Ferrari *et al.* 2017; Stevens *et al.* 2019). However, all of the ~65 species currently in culture are free-living, with the vast majority having been isolated from rotting fruits and flowers (Kiontke *et al.* 2011; Félix *et al.* 2014; Ferrari *et al.* 2017; Stevens *et al.* 2019).

Caenorhabditis bovis (Kreis 1964) is therefore particularly unusual for a *Caenorhabditis* species*,* having been isolated several times from the outer auditory canal of Zebu cattle in Eastern Africa (Kiontke & Sudhaus 2006) and recently from Gyr cattle in South America (Cardona *et al.* 2010). *C. bovis* is believed to be the causative agent of bovine parasitic otitis, a disease which causes a range of symptoms including inflammation, dark brown discharge from the affected ear, and dullness (Msolla *et al.* 1993). In severe cases, bovine parasitic otitis can result in mortality (Msolla *et al.* 1993). As is typical for a *Caenorhabditis* species, *C. bovis* is believed to have a phoretic association with an invertebrate, with larvae of the Old World screwworm fly (*Chrysomya bezziana*) also being found in the ears of Zebu cattle (Msolla *et al.* 1989, 1993). It is unclear to what extent bovine parasitic otitis is caused directly by *C. bovis* or by bacterial and/or fungal infections, and therefore to what extent *C. bovis* can be considered a parasite. Despite this, its close association with a vertebrate means that *C. bovis* is of particular interest to researchers interested in the evolution of nematode

parasitism and in *Caenorhabditis* diversity. However, as *C. bovis* is not in laboratory culture, it remains little studied.

In collaboration with local veterinarians and scientists, we sampled cattle at livestock markets and slaughterhouses in Western Kenya and successfully reisolated *C. bovis* from the ear of an adult female Zebu. We sequenced the genome of *C. bovis* in a nearby field laboratory using the Oxford Nanopore MinION platform and used the data to generate a high-quality, chromosome-scale draft genome sequence. We exploited this genome to determine the phylogenetic relationships of *C. bovis* to other species in the genus *Caenorhabditis,* including *C. elegans*, and reveal changes in genome size and content that have occurred during its evolution. We also reveal specific expansions in several gene families which may play a role in its unusual lifestyle. The high-quality draft genome and the analyses presented here represent a major step forward in our understanding of this unusual and understudied *Caenorhabditis* species.

Results

Reisolation of *C. bovis*

We sampled a total of 44 cattle of various ages and breeds at livestock markets and slaughterhouses in three counties in Western Kenya (Fig. 1A; Table S1). Sampling was performed by washing the outer auditory canal of each animal with cotton wool soaked in physiological saline (Fig. 1B) which was subsequently inspected under a dissecting microscope. We identified only a single instance of bovine parasitic otitis. The affected animal was an adult female Zebu which was sampled at a livestock market in Chwele, Bungoma County (Fig. 1A). The animal is believed to have originated from West Pokot County (Fig. 1A). Although we noted no obvious clinical symptoms, the cotton wool sample had an unpleasant odor, consistent with previous reports of bovine parasitic otitis (Msolla *et al.* 1993). We isolated approximately 50 live nematode larvae from the sample which were subsequently cultured on *E. coli*-seeded agar plates. The cultures thrived at 37°C on both nematode growth medium (NGM) and horse blood agar plates. Using a standard compound microscope, we identified adult nematodes as members of the genus *Caenorhabditis* based on their morphology (presence of a prominent pharyngeal bulb and filiform female tail). The morphology of the adult male tail (anteriorly closed fan, ray pattern with gap between GP2 and GP3, and a bent gubernaculum) was consistent with previous descriptions of *C. bovis* (Kreis 1964; Sudhaus & Kiontke 1996).

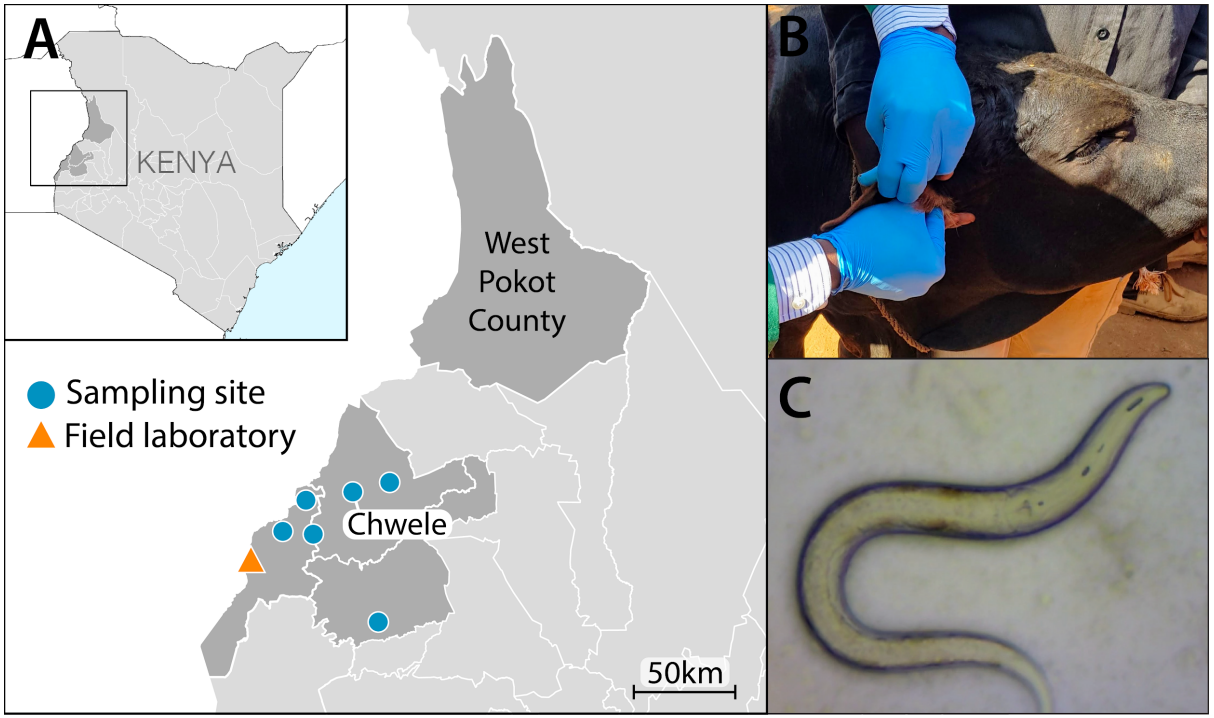

Figure 1: Cattle sampling and nematode isolation

A: Sampling locations in Western Kenya. We isolated *C. bovis* from an adult female Zebu sampled at a livestock market in Chwele. The animal was believed to have originated from West Pokot County. The location of the field laboratory in Busia is also shown. GPS coordinates and the number of animals sampled at each site can be found in Table S1. **B:** An animal being sampled using cotton wool soaked in physiological saline. **C:** Adult female *C. bovis* under a stereo microscope (*C. bovis* adults are ~1mm in length (Kreis 1964)).

A high-quality, chromosome-scale *C. bovis* reference genome

We sought to generate a high-quality reference genome for *C. bovis*. We took advantage of the portability of the Oxford Nanopore MinION platform and sequenced the genome of *C. bovis* in a field laboratory in Busia, Western Kenya (Fig. 1A). We generated 11.3 Gb of sequence data representing ~180-fold coverage of the *C. bovis* genome using two MinION v9.4 flow cells. Read length N50s were 11.4 kb and 4.3 kb, respectively, with the longest read spanning 242 kb (Table S2; Fig. S1). We also sequenced the genome to ~210-fold coverage (13.3 Gb) using the Illumina MiSeq platform at the BecA-ILRI Hub in Nairobi, Kenya. We identified and discarded reads originating from contaminant organisms, including several bacterial species which are known mammalian pathogens, using taxon-annotated GC-coverage plots (Fig. S2).

We assembled the *C. bovis* genome using the MinION long-reads and corrected residual sequencing errors in the assembly using the Illumina short-reads. The resulting assembly comprises 35 contigs spanning 62.7 Mb with a contig N50 of 7.6 Mb, with half of the assembly contained in just 4 contigs (Fig. 2A,B; Table 1). The assembly is highly complete, with 94.2% of a conserved set of nematode genes being present and fully assembled. In contrast to other outcrossing species, whose genomes typically contain highly levels of heterozygosity (Barrière *et al.* 2009), we find that the genome of *C. bovis* contains surprisingly little heterozygosity. Using a variant calling approach, we estimate that 0.03% of sites in the *C. bovis* genome are heterozygous (1 heterozygous site every ~3760 bp), with the k-mer distribution of the Illumina data indicating that the genome is essentially homozygous (Fig. S3). Using protein sequences predicted from the genomes of related nematodes as homology evidence, we predicted 13,128 protein-coding genes in the *C. bovis* genome. We note that this number is considerably lower than the number of genes predicted in the genomes of other *Caenorhabditis* species (Stevens *et al.* 2019). However, the gene set

contains 95.1% of a conserved set of nematode genes (Table 1), suggesting that the reduced count is not due to an incomplete gene set.

Chromosomal linkage groups are highly conserved in *Caenorhabditis* (Hillier et al. 2007). We defined 7,706 one-to-one orthologues between *C. bovis* and *C. elegans,* and exploited this conservation to assign 15 contigs (representing 99.4% of the *C. bovis* assembly) to the six *C. elegans* chromosomes (Fig. 2A; Fig. S4). Chromosomes III and V are represented by single contigs, suggesting that these contigs represent complete *C. bovis* chromosomes. Both contigs also show patterns of variation in GC content characteristic of the arms and centres organisation present in the chromosomes of other *Caenorhabditis* species (*C. elegans* Sequencing Consortium 1998; Hillier *et al.* 2007; Yin *et al.* 2018) (Fig. 2D,E). The remaining chromosomes are each represented by 3-4 contigs (Fig. 2A; Fig. S4).

	C. bovis v1	*C. elegans*
Span (Mb)	62.73	100.29
Number of contigs	35	7*
Contig N50 length (Mb)	7.56	17.49
Contig N50 number	4	3
Longest contig (Mb)	10.86	20.92
Repeat content (Mb)	8.19 (13.1%)	16.34 (16.3%)
BUSCO genome - complete (%) / fragmented (%)	94.2 / 4.6	98.7 / 0.7
Number of protein-coding genes	13,128	20,208
BUSCO gene set- complete (%) / fragmented (%)	95.2 / 3.2	98.7 / 0.9

Table 1: Genome and gene set metrics for *Caenorhabditis bovis* assembly v1.0. Assembly and gene set completeness was assessed using BUSCO (version 3.0.2) with the 'nematoda_odb9' dataset. *The *C. elegans* genome comprises 6 chromosomes and a 13 kb mitochondrial genome. WormBase ParaSite version WBPS12 of the *C. elegans* genome was used (Howe *et al.* 2017). MinION sequencing statistics are shown in Fig. S1 and Table S2. A taxon-annotated GC-coverage plot showing contigs from contaminating organisms that were removed from the *C. bovis* assembly is shown in Fig. S2. Kmer spectra of the Illumina short-read are shown in Fig. S3.

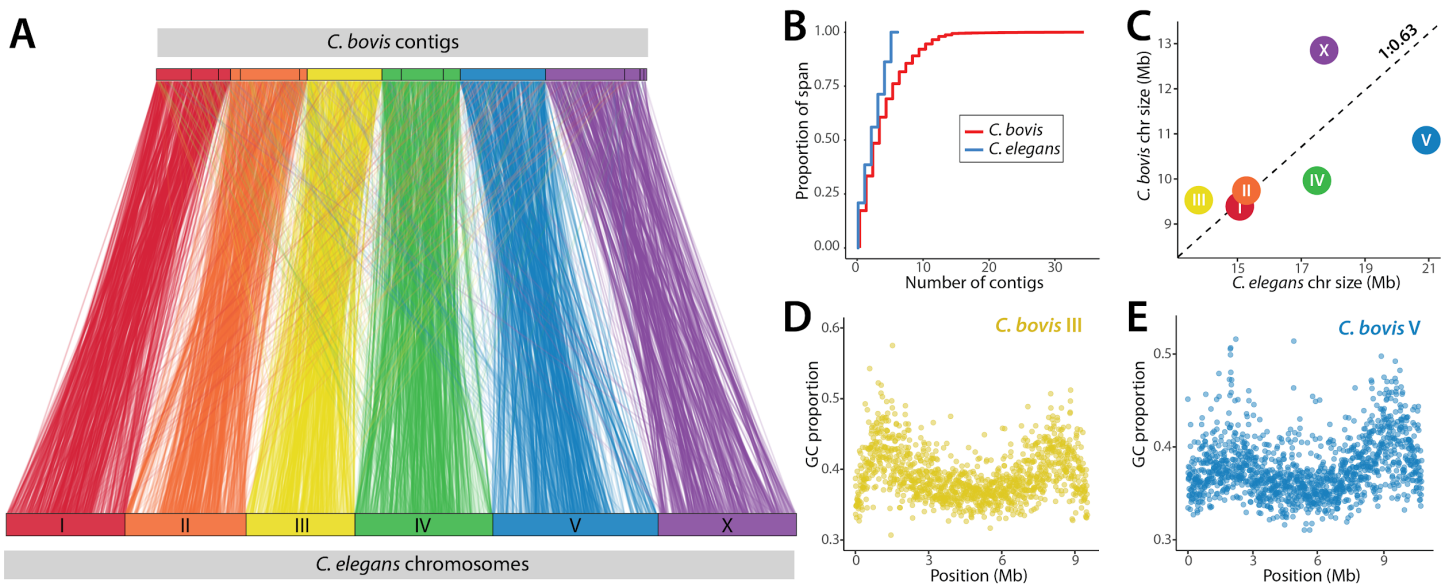

Figure 2: A high-quality, chromosome-scale *C. bovis* reference genome

A: Highly conserved linkage groups enable the assignment of 15 *C. bovis* contigs, comprising 99.4% of the assembly, to the six *C. elegans* chromosomes. Fig. S4 shows the genic composition of the 15 *C. bovis* contigs. Lines represent the position of 7,706 orthologues between *C. bovis* and *C. elegans*. **B:** Cumulative length as a proportion of span of the *C. bovis* and *C. elegans* genome assemblies. **C:** Chromosome size in *C. bovis* and *C. elegans*. Dotted line represents the expected chromosome size based on the proportion of overall genome size between *C. elegans* and *C. bovis* (1:0.63). **D, E:** Patterns of variation in GC content (using an 8 kb sliding window) in *C. bovis* contigs 3 (chromosome III) and 1 (chromosome V), respectively, are consistent with the arms and centers organisation present in the chromosomes of other *Caenorhabditis* species.

The position of *C. bovis* within *Caenorhabditis*

We sought to reconstruct the phylogenetic relationships of *C. bovis* to other species in the genus *Caenorhabditis*. We clustered over a million protein sequences predicted from the genomes of *C. bovis*, 32 other *Caenorhabditis* species (*C. elegans* Sequencing Consortium 1998; Stein *et al.* 2003; Mortazavi *et al.* 2010; Slos *et al.* 2017; Kanzaki *et al.* 2018; Yin *et al.* 2018; Stevens *et al.* 2019) and two outgroup taxa, *Diploscapter coronatus* (Hiraki et al. 2017) and *Diploscapter pachys* (Fradin et al. 2017), into orthologous groups and selected 1,167 single-copy orthologues. Alignments of these orthologues were concatenated to form a supermatrix which was used to reconstruct the *Caenorhabditis* phylogeny using maximum likelihood. Our phylogenomic analysis resulted in a well-supported phylogeny (Fig. 3) which was largely congruent with previously published phylogenies (Kiontke *et al.* 2011; Slos *et al.* 2017; Stevens *et al.* 2019). We recover *C. bovis* as sister to *Caenorhabditis plicata* with maximal support (bootstrap value of 100). The clade containing *C. bovis* and *C. plicata* is early-diverging within the genus *Caenorhabditis*, and the branches separating *C. bovis* and *C. plicata* are long, indicating that *C. bovis* is highly diverged from all other sequenced species, including *C. elegans*.

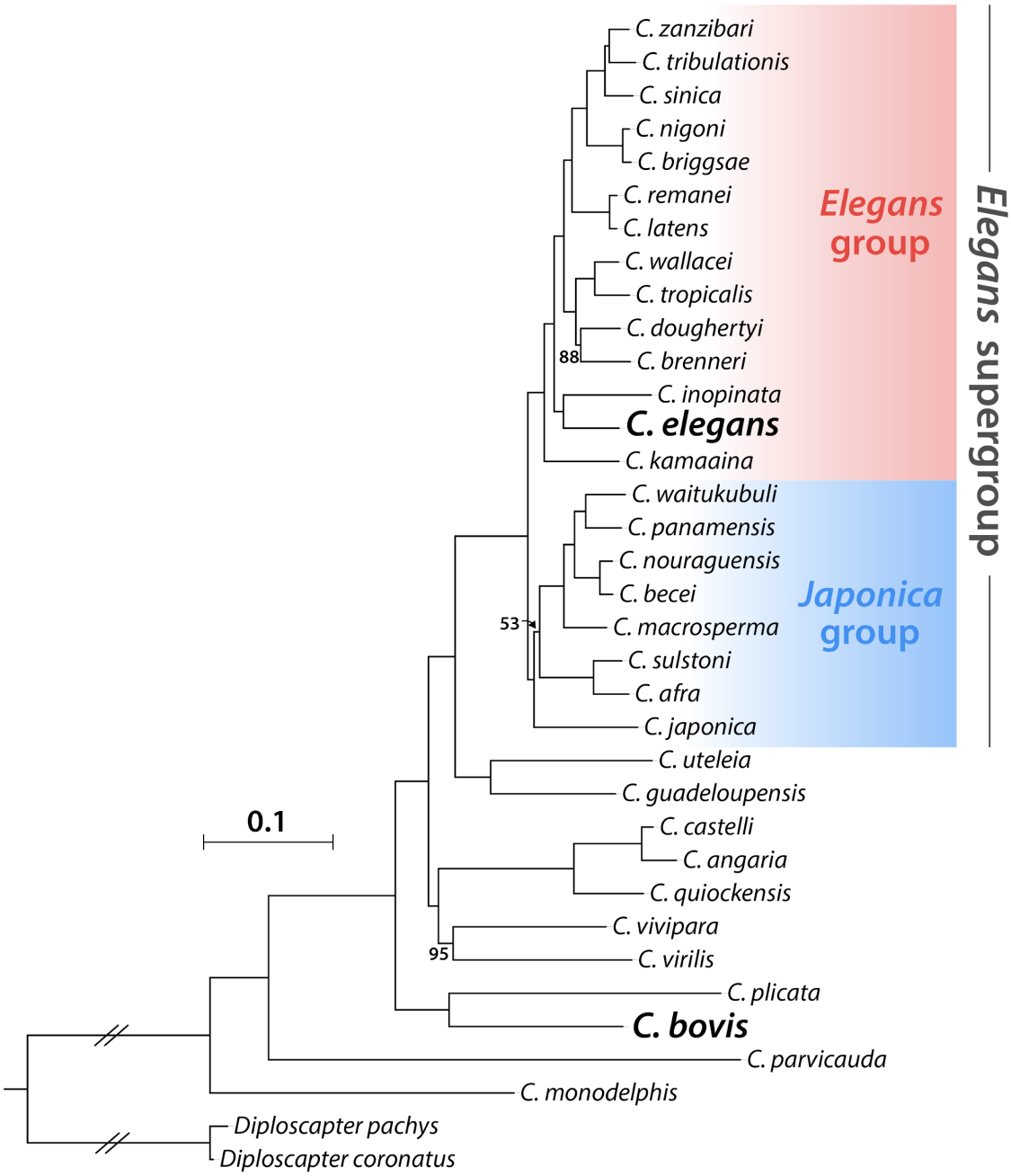

Figure 3: The phylogenetic position of *C. bovis* within *Caenorhabditis*

Phylogeny inferred using a supermatrix of 1,167 single-copy orthologues under the general time reversible substitution model with gamma-distributed rate variation among sites (GTR + Γ). *C. bovis* and *C. elegans* are highlighted in bold. The tree is rooted with the two *Diploscapter* species. Branch lengths are in substitutions per site; scale is shown. Bootstraps were 100 unless noted as branch annotations. Major clades as defined by (Kiontke *et al.* 2011) are highlighted.

Comparison between the *C. bovis* and *C. elegans* genomes

At 62.3 Mb, the *C. bovis* genome is the smallest *Caenorhabditis* genome published to date (Stevens *et al.* 2019), and is nearly 40 Mb smaller than the *C. elegans* genome. All six *C. bovis* chromosomes are smaller than their *C. elegans* homologues (Fig. 2C). Chromosome V is 49% smaller in *C. bovis* and contains fewer than half as many genes (2,302 and 4,992, respectively). Interestingly, the X chromosome is most conserved in size, being only 28% smaller in *C. bovis* and containing 20% fewer genes (2,260 and 2,782, respectively).

The majority of the overall difference in genome size (22.1 Mb or 58%) can be explained by a difference in protein-coding gene content, with the *C. bovis* genome containing 7,080 fewer predicted genes than *C. elegans*. The 13,128 *C. bovis* genes span 35.1 Mb, while the 20,208 *C. elegans* genes span 57.2 Mb, with genic DNA making up a similar proportion of each genome (56% and 57%, respectively). To understand what underlies the difference in gene number, we used the orthology clustering set described previously to compare the number of single-copy genes (those that do not cluster alongside another gene from the same species) and multi-copy genes (those that cluster alongside at least one other gene from the same species) in *C. bovis* and *C. elegans*. We find that the *C. bovis* gene set is substantially less redundant, with only 23% (3,095) of the gene set being classified as multi-copy, while 41% (8,351) of the *C. elegans* gene set is multi-copy. A particularly striking difference is in the number of G-protein coupled receptors (GPCRs), a large family of transmembrane proteins with chemosensory roles in *C. elegans (Robertson 1998)*. The *C. elegans* genome encodes 1,465 GPCRs, while the *C. bovis* genome contains just 326, accounting for 16% of the overall difference in gene number (Table S3). Several other large *C. elegans* gene families are also similarly underrepresented in the *C. bovis* genome, with differences in the number of nuclear hormone receptors (NHRs), major sperm proteins (MSPs), F-box proteins, and C-type lectins accounting for a further 10% of the difference in gene number (Table S3).

In addition to having fewer genes, the *C. bovis* genome contains a smaller proportion of repetitive DNA than the *C. elegans* genome (13% and 16%, respectively; Table S4), explaining a further 8.1 Mb (21%) of the difference in genome size. As is the case for *C. elegans*, repeats are underrepresented on the *C. bovis* X chromosome relative to the rest of the genome (6% versus 16%; Table S4). In *C. elegans*, repeats are distributed non-randomly within the five autosomes, with the chromosome arms being substantially more repeat rich than the centers (28% versus 9%; Fig. 4A,B; (*C. elegans* Sequencing Consortium 1998)). In contrast, we find a more even distribution of repeats in the two fully assembled *C. bovis* chromosomes (III and V), with the center regions being marginally more repeat-rich the arms (16% vs 14%, respectively, assuming the same proportional length of the arm and center domains as *C. elegans*; Fig. 5A,B).

We compared gene structure in 7,706 genes that were single-copy between *C. bovis* and *C. elegans*. Despite the genome being considerably smaller, *C. bovis* genes contain more introns than their *C. elegans* orthologues (8.6 and 6.6 introns per gene, respectively; Fig. 5C; Table S5). This is consistent with previous analyses which have found that early-diverging *Caenorhabditis* species have retained more ancestral introns than their in-group relatives (Kiontke *et al.* 2004; Slos *et al.* 2017). However, *C. bovis* introns are, on average, less than half the size of *C. elegans* introns (157 bp and 319 bp, respectively; Table S5). Therefore, despite containing more introns, *C. bovis* genes contain on average less intronic DNA than their *C. elegans* orthologues (1270 bp and 2375 bp of intronic DNA per gene, respectively; Fig. 5D; Table S5).

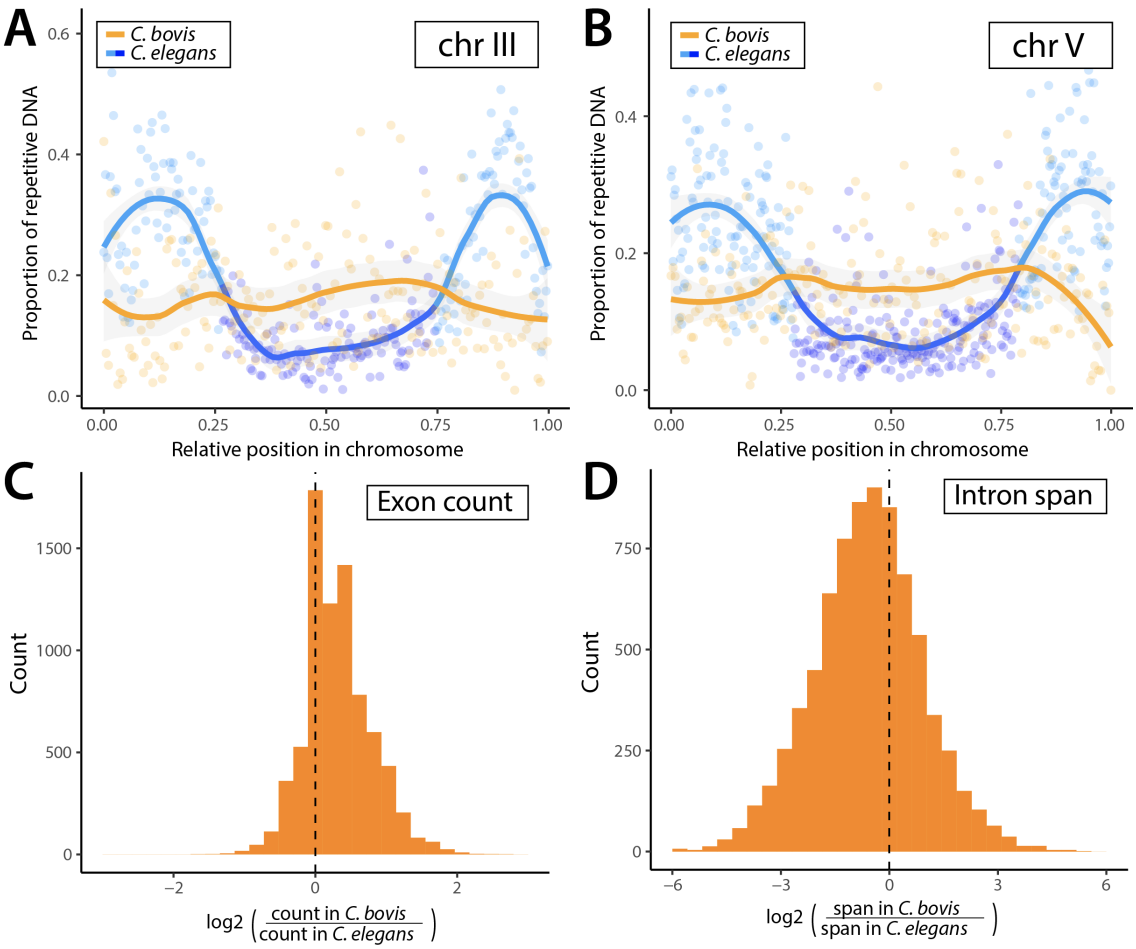

Figure 4: Comparison between the C. bovis and C. elegans genomes.

A, B: Repeat densities across chromosome III (**A**) and V (**B**) in 50 kb windows in *C. bovis* and *C. elegans*. Lines represent loess smoothing functions fitted to the data for each species. Points and lines for *C. elegans* are coloured by arms and centers domains (dark blue: centers, light blue: arms) as defined by (Rockman & Kruglyak 2009). Repeat content statistics for each chromosome are shown in Table S4. **C, D:** Histograms of the log2-transformed ratio of exon count (**C**) and intron span (**D**) in 7,706 genes in *C. bovis* compared to their orthologues in *C. elegans*. Untranslated regions (UTRs) are not annotated in *C. bovis* and so only coding exons and the intervening introns were considered in both species. Gene structure statistics are shown in Table S5. Counts of large gene families in *C. elegans* and *C. bovis* are shown in Table S3 and gene trees of expanded gene families in *C. bovis* are shown in Fig. S5.

Expanded gene families in the *C. bovis* genome

We sought to identify features of the *C. bovis* genome which may relate to its unusual ecology. Using the orthology clustering set described previously, we compared the *C. bovis* gene set to those of 32 other *Caenorhabditis* species. Despite having substantially fewer genes than other *Caenorhabditis* species, we identified several *C. bovis*-specific expansions in gene families which have independently been implicated in parasitism in other nematode species.

P-glycoproteins are members of the ATP-binding cassette (ABC) transporter family and are responsible for the removal of intracellular xenobiotics (Sheps *et al.* 2004). P-glycoproteins have been implicated in resistance to antihelminthic drugs in several parasitic nematode species (Xu *et al.* 1998; Bourguinat *et al.* 2008; Bartley *et al.* 2009). We find evidence for two duplications of the orthologue of the *C. elegans* P-glycoprotein gene *pgp-11* in *C. bovis*, resulting in three distinct copies (Fig. S5A). All other *Caenorhabditis* species, except for *C. monodelphis*, possess a single orthologue of *pgp-11* (Fig. S5A). *C. elegans* strains which lack *pgp-11* function show increased susceptibility to ivermectin (Janssen *et al.* 2013b), a widely used antihelminthic drug, and genetic variation in the orthologue of *pgp-11* is associated with variation in ivermectin susceptibility(Janssen *et al.* 2013a) in the horse parasite *Parascaris equorum (Janssen et al. 2013a)*.

Fatty acid and retinol (FAR) proteins are responsible for the uptake and transport of lipids required for nematode metabolism and development (Garofalo *et al.* 2003). FAR proteins have also been proposed to play a role in modulating host immune responses via the interference of lipid signalling pathways (Bradley *et al.* 2001). We find that the orthologue of C. *elegans* FAR gene, *far-8*, has undergone two duplications in *C. bovis*, resulting in three distinct copies. We also find evidence for the expansion of a family protein containing Kunitz-type serine protease inhibitor domains in *C. bovis*. A Kunitz-type serine protease inhibitor secreted by the hookworm *Ancylostoma ceylanicum* has been shown to be capable of inhibiting mammalian host proteases(Milstone *et al.* 2000). The majority of species, including *C. elegans,* possess

a single member of this family, while *C. bovis* possesses five. In addition, a family of galectin-domain containing proteins appears to be restricted to *C. bovis*. Galectins are actively secreted by several parasitic nematode species and may interfere with mammalian host immune responses(Turner *et al.* 2008; Kim *et al.* 2010; Wang *et al.* 2014).

Discussion

Here, we reisolated *C. bovis* from the ear of a female Zebu (*Bos taurus indicus*) in Western Kenya. We sequenced the genome of *C. bovis* using the Oxford Nanopore MinION platform in a nearby field laboratory and used the data to generate a high-quality, chromosome-scale reference genome. We exploited this genome sequence to reconstruct the phylogenetic relationships of *C. bovis* to other *Caenorhabditis* species, and identified expansions in gene families that may be associated with the unusual lifestyle of *C. bovis*.

The low level of heterozygosity in the *C. bovis* genome is surprising. Genomes of outcrossing *Caenorhabditis* species typically contain extremely high levels of heterozygosity (Cutter *et al.* 2006; Dey *et al.* 2013) which can complicate genome assembly (Barrière *et al.* 2009). To circumvent these issues, *Caenorhabditis* species are often deliberately inbred over multiple generations (e.g. by sibling mating) prior to sequencing (Stevens *et al.* 2019). While it is likely that our *C. bovis* cultures underwent some population bottlenecking during isolation and the subsequent two-week period of laboratory culture, if *C. bovis* has similar levels of heterozygosity to other outcrossing *Caenorhabditis* species, this alone is not sufficient to explain the low levels of heterozygosity we observe. Instead, it seems that the *C. bovis* population we sampled from is naturally highly inbred, suggesting that a very small number of nematodes are transported between hosts and that gene flow between demes is extremely rare. Resequencing other isolates would allow us to test if this is true of all *C. bovis* populations.

The placement of *C. bovis* as sister to *C. plicata* is intriguing. *C. plicata* has been isolated from carrion (once from a dead elephant in Kenya and once from a dead pine marten in Germany) and has a phoretic association with carrion beetles (Volk & Others 1950; Sudhaus 1974; Kiontke & Sudhaus 2006). *C. plicata* is therefore the only *Caenorhabditis* species currently in culture that has been found in association with a vertebrate, with all others having been isolated from rotting plant or fungal matter

(Kiontke *et al.* 2011; Félix *et al.* 2014; Ferrari *et al.* 2017; Stevens *et al.* 2019); Marie-Anne Felix, Lise Frezal, Matthew Rockman, Christian Braendle, John Wang, Michael Ailion, Erik Andersen, Asher Cutter, pers. comm.). In recent years, worldwide sampling has led to the discovery of many new *Caenorhabditis* species, but all efforts have been focussed in habitats that resemble the decaying vegetable matter habitat identified as the home of *C. elegans* (Félix & Braendle 2010; Kiontke et al. 2011; Félix & Duveau 2012; Félix et al. 2014; Ferrari et al. 2017). While there are anecdotal instances of other *Caenorhabditis* species being associated with vertebrates, including birds (Schmidt & Kuntz 1972), dogs (Kreis & Faust 1933), and humans (Scheiber 1880), no directed searches focussing on living or dead animal niches have been reported. It is therefore possible that there exists a largely undiscovered clade of vertebrate-associated *Caenorhabditis* species.

We do not yet know enough about the biology of *C. bovis,* from either its genome or its limited biological literature, to classify it as a 'true' parasite; *C. bovis* might instead be an opportunistic coloniser of niches created by other pathogens. Several other *Caenorhabditis* species associate with arthropods as phoretic hosts, and such phoresy is thought to serve as the major means of transport between scattered food patches (Kiontke & Sudhaus 2006). *C. bovis* is believed to be transported by dipterans (Msolla *et al.* 1989), themselves associated with parasitism of bovine ears, and may have exploited the biology of these phoretic hosts to colonise a new niche, the bovine ear. Bacterial coinfection may be a prerequisite of colonisation by *C. bovis*, may be exacerbated or encouraged by the presence of *C. bovis,* or may be initiated by *C. bovis* itself. Other Rhabditine species offer models for this last possibility: entomopathogenic species in the genus *Heterorhabditis* carry specific bacteria which play roles in killing their arthropod larval prey (Forst *et al.* 1997), and molluscicidal nematodes in the genus *Phasmarhabditis* induce bacterial sepsis in slugs and snail prey (Tan & Grewal 2002).

While the genome and the gene set of *C. bovis* is smaller than that of *C. elegans* and many other *Caenorhabditis* species*,* we identified several gene families that appear to have undergone expansion in *C. bovis*. Functional annotation of these expanded gene

families revealed that several have been independently implicated in parasitism in other nematode species. While P-glycoproteins (and orthologues of *pgp-11* specifically) are associated with resistance to antihelminthic drugs such as ivermectin, what role the expansion of *pgp-11* plays in the biology of *C. bovis* remains unclear. Ivermectin has been found to be effective at killing *C. bovis* and in the treatment of bovine parasitic otitis in cattle in Tanzania (Msolla *et al.* 1985). In contrast, similar studies have found ivermectin and albendazole to be an ineffective treatment for bovine parasitic otitis in cattle in Brazil (Verocai *et al.* 2009; Ferraz *et al.* 2019). Aside from P-glycoproteins, FAR proteins, galectins, and serpins are known to be actively secreted by several parasitic nematode species, and an immunomodulatory role has been proposed. It would be fascinating to explore the roles of these families (and many others) in the possible parasitic lifestyle of *C. bovis*. We note also that *C. bovis* appears to be adapted to life at 37°C in its bovine niche. This temperature is rapidly lethal to *C. elegans* (Snutch & Baillie 1983; Jones & Candido 1999) and thus *C. bovis* must have adapted to be heat resistant.

While the high-quality draft genome and the analyses presented here represent a major step forward in our understanding of this unusual and understudied *Caenorhabditis* species, it is only a beginning. Our ultimate goal is to establish long term cultures and to apply the exquisite reverse genetic toolkits available for *Caenorhabditis* to understand the biology of this species. We would like the isolates to be available to any researcher via the *Caenorhabditis* Genetics Center (CGC) and we are currently seeking the appropriate permits for export from Kenya. We hope that these cultures combined with the draft-genome sequence will enable the interrogation of the biology of *C. bovis*, including the use of CRISPR-Cas9 technology to edit or disrupt loci that might be relevant for its unusual lifestyle. It is important to note, however, that we still know very little about *C. bovis in situ*, with details of its present-day prevalence, role in bovine parasitic otitis and microbial associates remaining scarce. Therefore, any laboratory interrogation must happen alongside further study of *C. bovis* in Eastern Africa, in collaboration with local institutes and scientists.

Methods

Ethics Statement

This study was approved by the Institutional Research Ethics Committee (IREC Reference No. 2017-08) and the Institutional Animal Care and Use Committee (IACUC Reference No. 2017-04 and 2017-04.1) at the International Livestock Research Institute, review bodies approved by the Kenyan National Commission for Science, Technology and Innovation. Approval to conduct the work was also obtained from the Department of Veterinary Services and the relevant offices of these Ministries at the county government level. All recruited animal owners gave written, informed consent prior to their inclusion in the study.

Sampling, nematode isolation and culture

Sampling was carried out as part of an existing surveillance programme of zoonotic disease in humans at hospitals and livestock animals at livestock markets and slaughterhouses in three counties of Western Kenya (Fig. 1; Table S1). A total of 44 cattle, including a range of local breeds and ages, were sampled. We restrained each animal manually and washed the external auditory canal using cotton wool soaked in physiological saline. Cotton wool samples were stored in 50 ml tubes and transported to the laboratory in a refrigerated box. We inspected 1-2 ml of saline from each sample under a dissecting microscope within 4 hours of collection. Nematodes were isolated from the saline using a pipette and placed onto nematode growth medium (NGM) (1 g NaCl, 2 g Bactotryptone, 1.5 g KH2PO4, 0.25 g K2HPO4, 4mg cholesterol, 10g agar, 500 mL deionized water) or blood agar (50 mL horse blood, 41 g Columbia blood agar base, 1L of deionized water) plates seeded with an environmentally-sourced *E. coli* strain. Plates were incubated at 37°C. The morphology of adult nematodes was examined using a standard compound microscope and compared to the previous morphological descriptions of *C. bovis* (Kreis 1964; Sudhaus & Kiontke 1996).

DNA extraction

We harvested nematodes by washing each plate with phosphate-buffered saline (PBS) supplemented with 0.01% Tween20. The nematodes were washed three times with clean PBS and subsequently centrifuged to form a pellet. Pellets were stored at -40°C until extraction. We added 600 µL of Cell Lysis Solution (Qiagen) and 20 µL of proteinase K (20 µg/µL) to each frozen pellet and incubated for four hours at 56°C. 5 µL of RNase Cocktail Enzyme Mix (Invitrogen) was subsequently added and incubated at 37°C for one hour. We added 200 µl of Protein Precipitation Solution (Qiagen) and centrifuged at 15,000 rpm for 3 minutes. The supernatant was collected in a new tube and 600 µL of isopropanol added to precipitate the DNA. We centrifuged each tube at 15,000 rpm for 3 minutes and discarded the supernatant. The resulting DNA pellets were washed twice with 70% ethanol and briefly allowed to dry before being resuspended in 100 µL of elution buffer (10 mM Tris-Cl). DNA concentration was assessed using Qubit (Thermo Scientific).

Oxford Nanopore MinION sequencing

We sheared the DNA prior to sequencing by passing approximately 2 µg in a volume of 100 µl through either 26G or 29G insulin needle 5-10 times. Small fragments were removed by purifying DNA with 0.5x concentration Agencourt AMPure XP beads. We followed the "one-pot" ligation protocol for preparing Oxford Nanopore SQK-LSK108 libraries (https://www.protocols.io/view/one-pot-ligation-protocol-for-oxford-nanopore-libr-k9acz2e) but with the following modifications: we added 5 µl of SQK-LSK109 adapter mix (AMX) instead of 20 µl of SQK-LSK108 AMX; we added 20 µl of NEB Ultra II Ligation Master Mix instead of 40 µl; we replaced the SQK-LSK108 adapter binding beads (ABB) with either the SQK-LSK109 long fragment buffer (LFB) or short fragment buffer (SFB). Thereafter, we followed the standard manufacturer's instructions for preparing and loading SQK-LSK109 libraries. Libraries were loaded on to two R9.4 flow cell and run for ~48 hours using MinKNOW version 18.12.9. Raw data metrics are presented in Table S2.

Illumina MiSeq sequencing

We prepared one Nextera DNA Flex library as per manufacturer's instructions using ~100 ng of input DNA. The library fragment size was assessed using the Agilent TapeStation and library concentration was determined using Qubit dsDNA HS Assay Kit (Thermo Scientific, USA). The library was then sequenced using the Illumina MiSeq platform with a paired-end 300 bp MiSeq reagent kit v3 (Illumina Inc, USA) at the BecA-ILRI Hub in Nairobi, Kenya.

Genome assembly

Software versions and relevant parameters are available in the Zenodo repository. We base called the MinION FAST5 data using the high accuracy model in Guppy (available at https://community.nanoporetech.com). We generated a preliminary assembly using wtdbg2 (Ruan & Li 2019) and identified contaminants using taxon-annotated, GC-coverage plots (Fig. S2) as implemented in blobtools (Laetsch & Blaxter 2017a). Reads were mapped to the preliminary assembly using minimap2 (Li 2018) and the likely taxonomic origin of each contig was determined by searching NCBI nucleotide 'nt' or UniProt Reference Proteomes (Pundir *et al.* 2017) using NCBI-BLAST+ (Camacho *et al.* 2009) or DIAMOND (Buchfink *et al.* 2015), respectively. Reads originating from contaminant organisms were discarded. We generated the final assembly using wtdbg2. Sequencing errors were initially corrected by aligning the MinION reads to the assembly using minimap2 and performing four iterations of Racon (Vaser *et al.* 2017) followed by a single iteration of Medaka (available at https://github.com/nanoporetech/medaka). Any remaining errors were corrected by aligning the Illumina MiSeq reads to the assembly using BWA-MEM (Li 2013) and performing two iterations of Racon followed by two iterations of Pilon (Walker *et al.* 2014).

Gene prediction

Prior to gene prediction, repeat sequences were identified *de novo* using RepeatModeler (Smit & Hubley 2010) and subsequently masked using RepeatMasker (Smit *et al.* 1996). Protein-coding genes were predicted using BRAKER (Hoff *et al.*

2016), using proteins sequences from nematode-specific EggNOG database (which comprises sequences from *C. elegans*, *C. briggsae*, *C. remanei*, *C. japonica, Pristionchus pacificus* and *Trichinella spiralis*) (Huerta-Cepas *et al.* 2016b) as homology evidence. Genome assembly and gene set completeness were assessed using BUSCO with the 'nematoda_odb9' database (Simão *et al.* 2015).

Estimation of heterozygosity

We estimated heterozygosity in the *C. bovis* genome using two approaches. We used Jellyfish (Marçais & Kingsford 2011) to count kmers (k=19) in adapter-trimmed and contaminant-free Illumina MiSeq reads and used the GenomeScope website (Vurture *et al.* 2017) to estimate heterozygosity. To specifically call heterozygous sites in the *C. bovis* genome, we aligned the Illumina MiSeq reads to the *C. bovis* assembly using BWA-MEM and removed possible PCR duplicates from the resulting BAM file using PicardTools ("Picard Tools - By Broad Institute" 2019). We performed variant calling using freebayes (Garrison & Marth 2012) and used bcftools (Danecek *et al.* 2014) to remove variants sites that were dependent on strand or the position of the aligned read. We then estimated heterozygosity by dividing the total number of biallelic single nucleotide polymorphisms (SNPs) by the total number of sites (only those sites with a read depth $\geqslant 8$ and $\leqslant 250$, which represented 99.3% of the genome, were considered).

Assignment of *C. bovis* contigs to chromosomes

To assign *C. bovis* contigs to chromosomes, we identified one-to-one orthologues between *C. bovis* and *C. elegans* using a reciprocal best BLAST hit approach. Both proteomes were filtered so that they contained only the longest-isoform per gene and searched against each other using blastp. Protein pairs which had reciprocal best BLAST hits with e-values < 1e-25 and a query coverage >75% were declared as one-to-one orthologues. *C. bovis* contigs containing 10 or more *C. elegans* orthologues were assigned to the chromosome containing the majority of the *C. elegans* orthologues.

Orthology inference and phylogenomics

Accession details for all data used in this analysis are available in the Zenodo repository. We selected the protein sequence of the longest isoform of each protein-coding gene in *C. bovis*, 32 other species of *Caenorhabditis,* and the two outgroup taxa, *Diploscapter coronatus* and *Diploscapter pachys*. OrthoFinder (Emms & Kelly 2015) was used to cluster all protein sequences into putatively orthologous groups (OGs) using the default inflation value of 1.5. OGs containing loci which were present in at least 75% of species and which were, on average, single copy (mean count per species < 1.3) were selected. We aligned each selected OG using MAFFT (Katoh & Standley 2013) and generated a maximum likelihood tree along with 1000 ultrafast bootstraps (Hoang *et al.* 2018) using IQ-TREE (Nguyen *et al.* 2015), allowing the best-fitting substitution model to be selected automatically (Kalyaanamoorthy *et al.* 2017). Each tree was screened by PhyloTreePruner (Kocot *et al.* 2013), collapsing nodes with bootstrap support <90, and any OGs containing paralogues were discarded. If two representative sequences were present for any species (i.e., "in-paralogues") after this paralogue screening step, only the longest of the two sequences was retained. We then realigned the remaining OGs using MAFFT and trimmed spuriously aligned regions using trimAl (Capella-Gutiérrez *et al.* 2009). The trimmed alignments were subsequently concatenated to form a supermatrix using catfasta2phyml (available at https://github.com/nylander/catfasta2phyml). We inferred the species tree using IQ-TREE with the general time reversible model (GTR) with gamma-distributed rate variation among sites. The resulting tree was visualized using the iTOL web server (Letunic & Bork 2016)).

Gene content and structure analyses

To understand the large difference in protein-coding gene number between *C. bovis* and *C. elegans*, we used the orthology clustering set described previously to determine the level of redundancy in each gene set. For each species, we counted the number of loci in orthogroups containing two or more representatives from that species (multi-copy) and the number of loci in orthogroups containing a single representative from that species (single-copy). We also searched the longest isoform

of each protein-coding gene from both species against the Pfam (Bateman *et al.* 2004) database using InterProScan (Jones *et al.* 2014). We then counted the number of loci in each species that were annotated as being GPCRs, NHRs, MSPs, F-box proteins, or C-type lectins. These gene families are known to constitute a substantial fraction of the *C. elegans* gene set ("Genomic classification of protein-coding gene families" 2019).

We sought to identify gene families that have undergone expansion in the relatively small *C. bovis* gene set. We provided the orthology clustering set described previously to KinFin (Laetsch & Blaxter 2017b) to compare counts between *C. bovis* and all other species. We also searched the longest isoform of each protein-coding gene for all species against Pfam using InterProScan and provided the output to KinFin to annotate each orthogroup with a putative function. We screened the expanded gene families for functions that had previously been implicated in parasitism in other nematode species. To further affirm expansion in *C. bovis,* we generated gene trees for each orthogroup of interested using IQ-TREE as previously described.

To compare gene structure in *C. bovis* and *C. elegans*, we identified one-to-one orthologues in the orthology clustering set and extracted the exon counts and intron spans from the GFF annotation files of each species. As untranslated regions (UTRs) are not annotated in *C. bovis*, only coding exons and intervening introns were considered for both species. We calculated log2-transformed ratios of exon counts and intron spans for each gene pair using a Python script (available at https://github.com/lstevens17/cbovis_manuscript).

Repeat content analyses

To generate comprehensive repeat libraries and annotations for both *C. bovis* and *C. elegans*, we followed the approach of (Berriman *et al.* 2018). Briefly, we used TransposonPSI (Haas 2007) to identify transposon sequences in both species, retaining those that were at least 50 bp in length. We also identified long terminal repeat (LTR) transposons in each species using LTRharvest (Ellinghaus *et al.* 2008).

We searched the resulting library with protein HMMs from Pfam and GyDB (Llorens *et al.* 2011) using LTRdigest (Steinbiss *et al.* 2009) and discarded any sequence that did not contain a transposable element domain. We also identified repetitive sequences *de novo* in both species using RepeatModeler. We then combined the resulting repeat libraries, classified each sequence using RepeatClassifier, and clustered the libraries at an identity of ⩾80% using VSEARCH (Rognes *et al.* 2016) to create a non-redundant repeat library for each species. We removed sequences from these repeat libraries that had significant homology to any member of the *C. elegans* gene set using TBLASTN. The resulting non-redundant and filtered repeat libraries were then provided to RepeatMasker which generated the final repeat annotations for each species along with genome file with repeat sequences masked with N's. We used a Python script (available at https://github.com/lstevens17/cbovis_manuscript) to compute repeat densities in 50 kb windows across chromosomes III and V in both *C. bovis* and *C. elegans*.

Quantification and statistical analysis

Statistical analyses were conducted using R (v3.5.1) (Team 2018) and Python 2.7. Gene structure ratios were log2-transformed using the math Python module. Loess smoothing curves were fitted to repeat densities using the ggplot2 R package (v2.3.1) (Wickham 2009).

Data and code availability

Raw sequence data and the genome assembly and annotation files have been deposited in the relevant INSDC databases under the accession PRJEB34497. The assembly and gene set are also available to browse, query, and download at http://www.caenorhabditis.org. Data files associated with this study have been deposited in Zenodo under the accession 10.5281/zenodo.3571457. Scripts and intermediate files associated with this study are available in the following GitHub directory https://github.com/lstevens17/cbovis_manuscript.

Chapter 4

The phylogeny of the genus *Caenorhabditis*

Abstract

Caenorhabditis elegans is a key laboratory model organism. An improved understanding of the natural ecology of *C. elegans*, combined with worldwide sampling efforts, has led to the discovery of many new species of *Caenorhabditis*. A genus-wide genome sequencing project has generated draft genomes for many of these new species. However, before they can be exploited for evolutionary study, an understanding of the phylogenetic relationships in the genus is required. Here, I use the genomes and transcriptomes of 58 *Caenorhabditis* species and two outgroup taxa to perform the most comprehensive reconstruction of the *Caenorhabditis* phylogeny to date. The resulting topologies are well-resolved and well-supported and reaffirm the monophyly subgeneric groups recovered by previous analysis, including the *Elegans* supergroup. However, I find disagreements with previously published phylogenies, including recovering the *Drosophilae* supergroup as paraphyletic. I also find two relationships that are inconsistent across our analyses and further investigate the origins of these conflicts. Our results provide a fundamental phylogenetic framework for future evolutionary studies within this important genus.

Introduction

Caenorhabditis elegans has become one of the preeminent model organisms in modern biology, but only recently have we started to understand its evolutionary history (Félix & Braendle 2010). An improved understanding of the natural ecology of *C. elegans*, combined with worldwide sampling efforts, has led to the discovery of many new species of *Caenorhabditis*, with over 60 species currently in laboratory culture (Kiontke *et al.* 2011; Félix *et al.* 2014; Ferrari *et al.* 2017; Stevens *et al.* 2019); Marie-Anne Félix, Lise Frezal, Matthew Rockman, Christian Braendle, John Wang, Michael Ailion, Erik Andersen, Asher Cutter, pers. comm.). A genus-wide genome sequencing project has generated draft genomes for over 50 of these species (discussed in Chapter 2), and these data promise to provide an essential evolutionary context for *C. elegans* and vast body of associated research. However, before these new species and their genomic resources can be exploited for comparative study, an understanding of the phylogenetic relationships within the genus is required.

Early reconstructions of the phylogeny of the genus *Caenorhabditis* were performed using morphological characters, including several features of the male mating apparatus (Sudhaus & Kiontke 1996). As DNA sequencing became commonplace, phylogenetic analyses using small numbers of nuclear loci were conducted (Fitch *et al.* 1995; Cho *et al.* 2004; Kiontke *et al.* 2004, 2011). These analyses defined major subgeneric clades of species: the *Elegans* supergroup, which contains the *Japonica* and *Elegans* groups, and the *Drosophilae* supergroup, which contains the *Drosophilae* and *Angaria* groups (Kiontke *et al.* 2011). These molecular-based analyses also revealed that many of the characters previously used for phylogenetic inference, including reproductive mode, have evolved multiple times independently reducing their suitability for inferring phylogenetic relationships within the genus. Recently, phylogenies based on draft genome sequences, which have exploited far larger amounts of data, have been published (Slos *et al.* 2017; Stevens *et al.* 2019). These new phylogenomic analyses have challenged the results of previous analyses, particularly the monophyly of the *Drosophilae* supergroup.

There is substantial disagreement in the field of phylogenetics surrounding the most appropriate way to analyse large, multi-locus datasets (Jeffroy *et al.* 2006). The conventional approach involves concatenating alignments of individual loci into a single alignment, known as a supermatrix, which is subsequently used to infer the species tree. The large number of sites in these supermatrices enable the use of complex substitution models that may more accurately reflect the substitution process (Lartillot & Philippe 2004). However, this approach is known to lead to inaccurate topologies when high levels of conflicting signals (e.g. due to incomplete lineage sorting (ILS) or introgression) are present in the dataset (Kubatko & Degnan 2007). ILS related to effective population size and is therefore expected to be particularly problematic in lineages with high effective population, such as *Caenorhabditis*. An alternative approach, known as the supertree or summary approach, is to infer trees independently for subsets of the data in close linkage (usually each locus) and use a separate tool to infer the most likely species tree (Sanderson *et al.* 1998). While many of these tools have been specifically designed to deal with conflicting signals (Zhang *et al.* 2018), they are sensitive to errors in gene trees (Roch & Warnow 2015), which are extremely common in datasets containing many species (Gatesy & Springer 2014). A third approach, known as the multi-species coalescent approach, involves co-estimation of the gene trees and species tree (Heled & Drummond 2010). While in theory this approach circumvents the major limitations of both the supermatrix and supertree approaches, it is not yet computationally tractable to perform multi-species coalescent analyses on datasets containing many species and/or loci and, as a result, they are not widely used (Zimmermann *et al.* 2014).

Here, I use the genomes and transcriptomes of 58 *Caenorhabditis* species and two outgroup taxa to perform the most comprehensive reconstruction of the phylogeny of the genus *Caenorhabditis* to date. The recovered topologies, inferred using over 2,000 single-copy orthologues, are well-resolved and well-supported and reaffirm the monophyly of the *Elegans* supergroup and groups therein. I also recover the

monophyly of the *Angaria* and *Drosophilae* groups, and propose names for two new clades. However, I find disagreements with previously published phylogenies, including recovering the *Drosophilae* supergroup as paraphyletic. I also find two relationships that are inconsistent across our analyses and further investigate the origins of these conflicts. My analyses provide a fundamental phylogenetic framework for future evolutionary studies within this important genus of nematodes.

Results

The *Caenorhabditis* phylogeny

I performed orthology clustering of 1,357,355 protein sequences predicted from the genomes and transcriptomes of 58 *Caenorhabditis* species and two outgroup taxa, *Diploscapter coronatus* and *Diploscapter pachys*. I identified 2,869 single-copy orthologues, each of which was present in at least 45 of the 60 taxa, and aligned their amino acid sequences. I employed three approaches to reconstruct the species tree. First, I concatenated the alignments of all 2,869 single-copy orthologues into a supermatrix containing 821,243 sites and estimated the species tree using maximum likelihood (ML) under the general time reversible model (GTR). I also estimated the species tree using Bayesian inference (BI) under the more complex CAT-GTR model, which accounts for among-site variation in substitution processes, using a smaller supermatrix consisting of 467 single-copy orthologues containing 185,051 sites. Lastly, I employed a supertree approach by estimating gene trees for all 2,869 single-copy orthologues and providing the resulting topologies to ASTRAL-III to estimate the species tree.

The three analyses yielded highly congruent, well-supported topologies that displayed very few inconsistencies, discussed below (Fig. 1; Fig. S1–3). The majority of relationships, including the monophyly of the *Elegans* supergroup and of the *Angaria, Drosophilae*, *Elegans*, and *Japonica* groups, as defined by (Kiontke *et al.* 2011), were recovered with maximal support by all analyses. I recovered a clade containing *C. guadeloupensis*, *C.* sp. 45, and *C. uteleia* (which I designate as the *Guadeloupensis* group) as sister to the *Elegans* supergroup, and therefore find the *Drosophilae* supergroup, as defined by (Kiontke *et al.* 2011), to be paraphyletic. I also recover a clade containing *C. portoensis, C. vivipara,* and *C.* sp. 27 which I name the *Portoensis* group. These taxa were not considered by (Kiontke *et al.* 2011). A majority of species are members of the *Elegans* supergroup, which comprises 35 of the 58 sequenced *Caenorhabditis* species. *C. monodelphis* is the earliest diverging species and

the node joining *C. monodelphis* with the remaining taxa defines the genus *Caenorhabditis*.

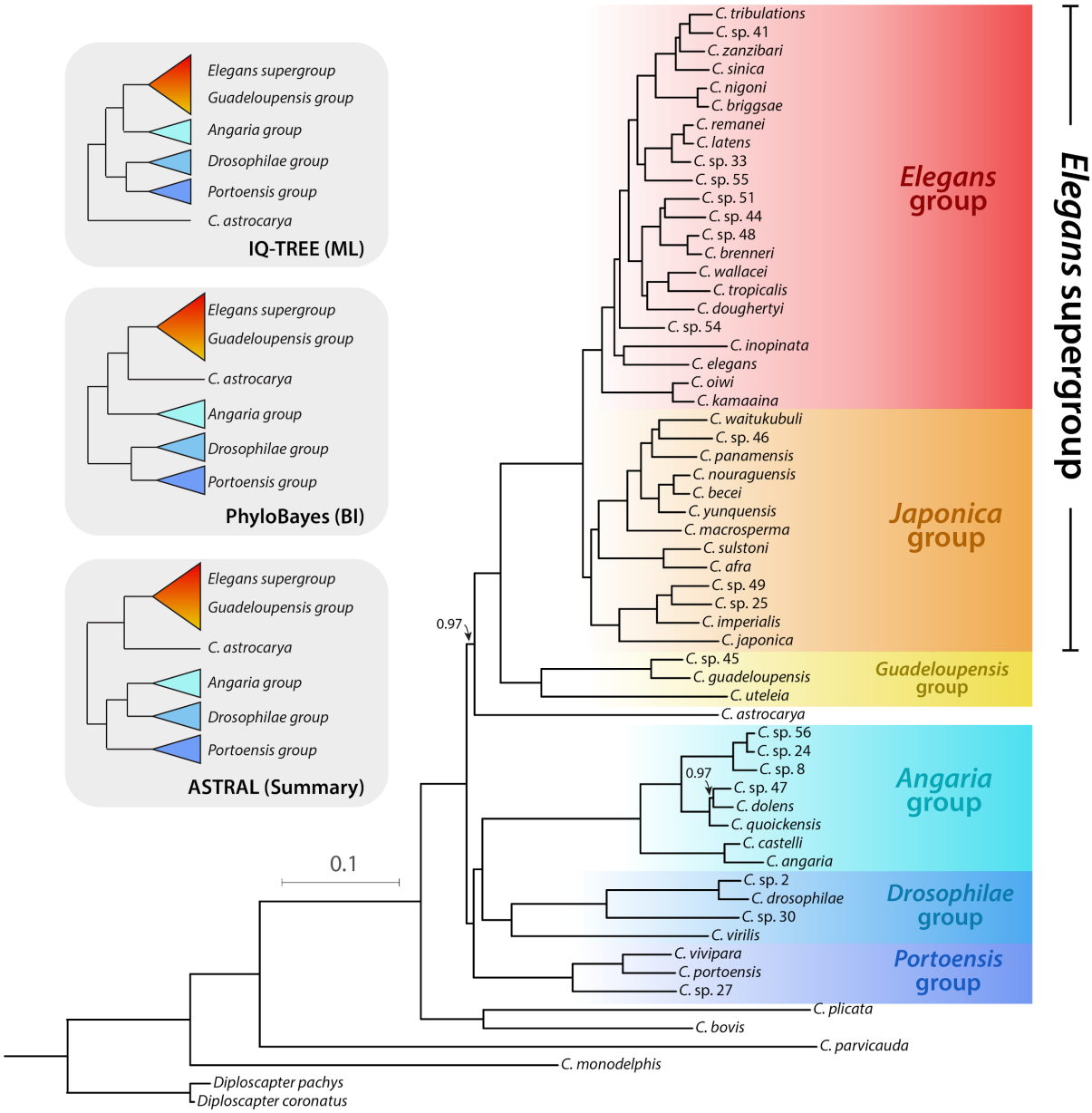

Figure 1: Phylogenetic relationships within the genus *Caenorhabditis*.

Phylogeny inferred using 2,869 gene tree with ASTRAL-III. Bayesian posterior probabilities were 1 unless noted as branch annotations. Branch lengths in substitutions per site were estimated using IQ-TREE (GTR+Γ); scale is shown. Major clades are highlighted. Inset: alternative topologies recovered by each analysis.

Two contentious relationships

Interestingly, two relationships were inconsistent across my analyses. Both the BI and supertree approach recovered *C. astrocarya* as sister to the clade containing the *Elegans* supergroup and the *Guadeloupensis* group*,* while ML recovered *C. astrocarya* as early diverging within the genus (Fig. 1). Secondly, the supertree approach recovered the *Angaria Drosophilae,* and *Portoensis* groups as a single monophyletic group, while both concatenation approaches (ML and BI) recovered this group as paraphyletic, with the *Angaria* group as more closely related to the clade containing the *Elegans* supergroup and *Guadeloupensis* group (Fig. 1). Despite representing conflicting phylogenetic hypotheses, these relationships were often recovered with maximal support (bootstrap values of 100 and Bayesian posterior probabilities of 1). I note that all conflicting branches occur deep within the phylogeny and are very short (Fig. 1; Fig S1-3; Table S1).

To explore support for these contentious relationships, I calculated gene concordance factors (gCFs) (Minh *et al.* 2018) by determining the proportion of gene trees that were concordant with each internal branch in all threerecovered topologies. While the proportion of the 2,869 gene trees that were concordant with the undisputed branches was highly variable between branches (20.8-99.8%), an extremely low proportion of gene trees were concordant with each of the contentious branches (4.6-10.7%) (Fig. S1-3). The placement of *C. astrocarya* as sister to the clade containing the *Elegans* supergroup and *Guadeloupensis* group was recovered in marginally more gene trees than the alternative placement (10.7% and 6.3%, respectively). The monophyly of the clade containing *Angaria, Drosophilae* and *Portoensis* groups was recovered in 5.6% of gene trees, while the branches supporting the paraphyly of this clade were recovered in 4.7% and 6.4% of gene trees (Fig. S1-3). As all contentious branches are short and occur relatively early in the diversification, I sought to assess whether this explained the high level of discordance among gene trees. Using multiple linear regression, I found that the majority of the variance in gene tree concordance in each branch is explained by branch length and distance from the root ($P<0.001$; $r2=0.70$; Table S1). This suggests that the inconsistency of

these relationships across analyses and the high level of incongruence among gene trees largely arise from limited phylogenetic signal, rather than processes such as ILS or introgression.

Given that this limited phylogenetic signal results in a high number of gene trees that are discordant with any of the plausible hypotheses, I instead opted to assess gene-wise support by performing constrained gene tree searches. Using 581 single-copy orthologues present in all species, I performed constrained gene tree searches under each phylogenetic hypothesis (Fig 2) and assessed support using the approximately unbiased (AU) test (Shimodaira 2002). The placement of *C. astrocarya* as sister to the clade containing the *Elegans* supergroup and the *Guadeloupensis* group was supported with the highest probability by 315 loci, while 265 loci supported the alternative hypothesis (Fig. 2A). A single monophyletic clade comprising the *Angaria, Drosophilae,* and *Portoensis* groups was supported with the highest probability by 323 loci, while 257 loci supported the paraphyly of this group. However, in both analyses, an extremely low proportion of loci supported either hypothesis significantly ($P < 0.05$) better than the alternative (Fig. 2), reaffirming my finding that there is limited phylogenetic signal to resolve these relationships. The two most well-supported relationships were recovered by the supertree approach and this topology therefore represents my best estimate of the *Caenorhabditis* phylogeny (Fig. 1).

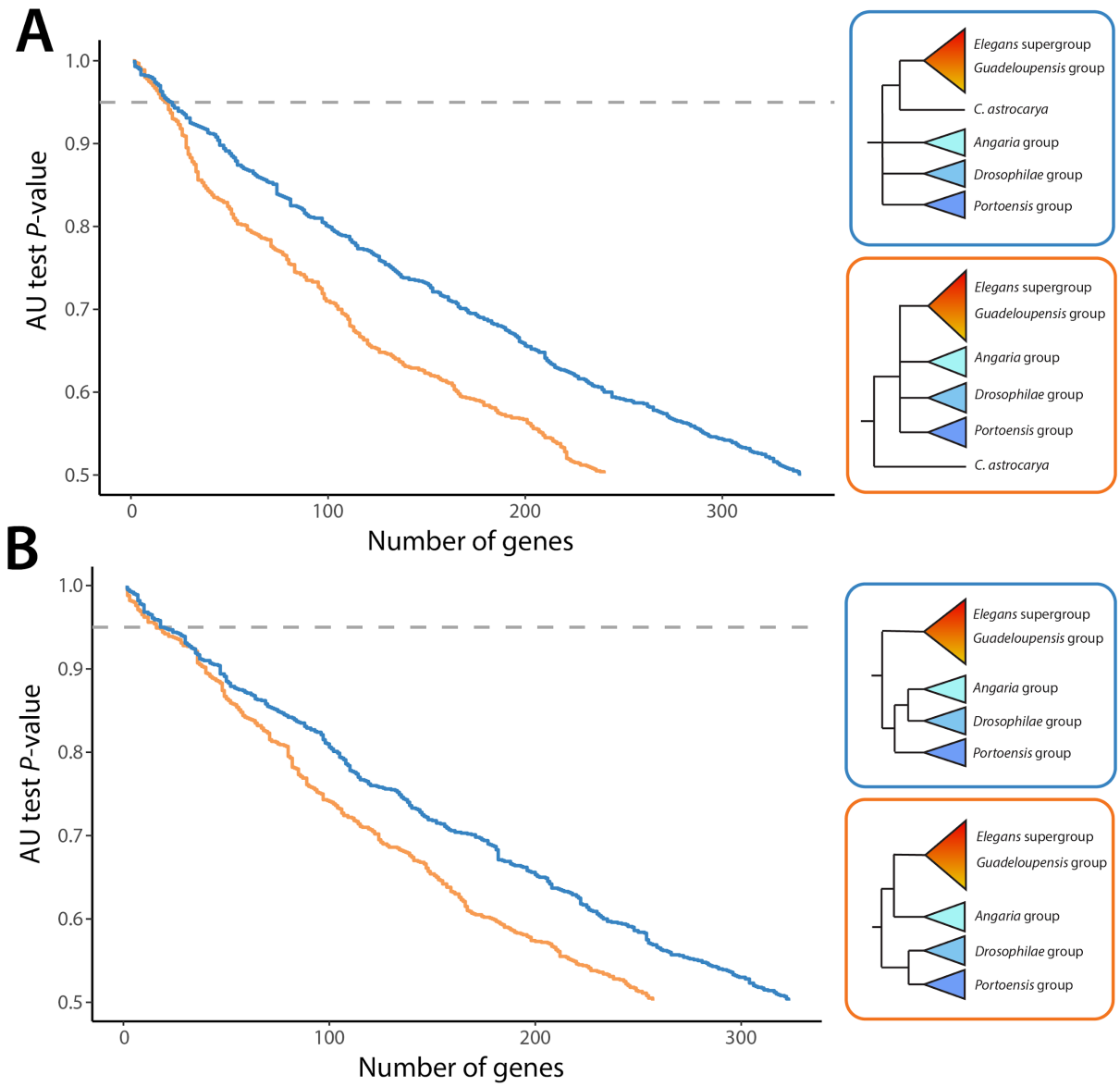

Figure 2: Gene-level support for competing phylogenetic hypotheses.

Constrained gene tree searches were conducted for 581 single-copy orthologues under each hypothesis. Lines represent the cumulative number of genes that most strongly support each hypothesis and their associated *P*-values. Gene trees above the dashed line were significantly more well-supported than the alternative hypothesis ($P < 0.05$). **A:** Support for two competing hypotheses concerning the placement of *C. astrocarya*. The relationships between the *Angaria*, *Drosophilae,* and *Portoensis* groups were unconstrained. **B:** Support for two competing hypotheses concerning the relationships between the *Angaria*, *Drosophilae,* and *Portoensis* groups. The position of *C. astrocarya* was unconstrained.

Discussion

Using genomic and transcriptomic data from 58 species of *Caenorhabditis* and two outgroup taxa from the genus *Diploscapter*, I have conducted the most comprehensive analysis of the *Caenorhabditis* phylogeny to date. The majority of relationships were recovered with maximal support by all analyses and are consistent with previous studies (Kiontke *et al.* 2011; Slos *et al.* 2017; Stevens *et al.* 2019). Our results corroborate the monophyly of the *Japonica* and *Elegans* groups and of the *Elegans* supergroup as defined by (Kiontke *et al.* 2011). I also find the *Angaria* and *Drosophilae* groups to be monophyletic and propose names for two further monophyletic groups: the *Portoensis* group and the *Guadeloupensis* group. My finding that the *Guadeloupensis* group is sister to the *Elegans* supergroup is consistent with previous phylogenomic analyses (Slos *et al.* 2017; Stevens *et al.* 2019) and provides further evidence that the *Drosophilae* supergroup, as defined by (Kiontke *et al.* 2011), is paraphyletic.

My finding that two relationships were inconsistent across analyses adds to a growing body of findings that the topologies recovered by large phylogenomic analyses, despite frequently receiving unequivocal support, are often highly dependent on dataset, taxon sampling and methodology (Jeffroy *et al.* 2006). I find that gene-tree concordance factors (Minh *et al.* 2018) reveal variation in support that is otherwise obscured by traditional bootstrap and Bayesian support values and therefore represent a more suitable measure of support in large phylogenomic datasets. However, the extremely low number of gene trees that were concordant with the alternative topologies suggests that a high degree of gene tree estimation error is present in our dataset. In order to determine which of a set of plausible hypotheses is most probable, constrained gene tree searches and topology tests, such as those performed by ourselves and others (Arcila *et al.* 2017), may be more suitable when gene-tree estimation error is pervasive.

It is interesting that the most well-supported topology was recovered by the supertree approach. Coalescent-based supertree approaches, such as ASTRAL-III (Zhang *et al.* 2018), are specifically designed to accommodate conflicting signals in gene trees arising from ILS. ILS is expected to be particularly common in rapidly diverging populations with large effective population sizes. The conflicting bipartitions in our analyses all had short branches and outcrossing *Caenorhabdtis* species are known to have extremely large effective population sizes (Cutter *et al.* 2006; Dey *et al.* 2013). It is therefore possible that ILS is leading to conflicting signals at these branches that misleads our concatenation-based approaches. However, several other branches in the phylogeny are of similar length but are recovered by both concatenation and summary approaches, suggesting that conflicting signals due ILS are not the primary cause of the inconsistencies I observe. Instead, it appears that, despite the size of our dataset, there is insufficient phylogenetic signal to resolve relationships between species which diverged rapidly and early in the evolution of the genus.

Despite these remaining questions, the phylogeny of *Caenorhabditis I* present is largely fully supported. The relationships I present will form the essential backbone of future comparative phylogenomic analyses, placing the exquisitely well-understood biology of *C. elegans* in a rich evolutionary context. Future work could explore additional data types, such as rare genomic changes (e.g. gain and loss of genes and other features, or clade-restricted insertion-deletion events) to fully resolve the phylogeny (Rokas & Holland 2000).

Methods

Details of software versions and parameters used in these analyses are available in Table S2.

Orthology Inference and single-copy ortholog selection

Details of all data used in the orthology clustering analysis are available in Table S3. I collected the proteins sequences predicted from the genomes and transcriptomes of 58 *Caenorhabditis* species and two outgroup taxa (*Diploscapter coronatus* and *Diploscapter pachys*) and selected the longest isoform of each gene. OrthoFinder (Emms & Kelly 2015) was used to cluster all protein sequences into putatively orthologous groups (OGs). OGs which were, on average, single-copy and present in at least 75% of species were selected using KinFin (Laetsch & Blaxter 2017b). To identify paralogous sequences, I aligned the protein sequences of each selected OG using MAFFT (Katoh & Standley 2013) and generated a maximum likelihood tree along with 1000 ultrafast bootstraps (Hoang *et al.* 2018) using IQ-TREE (Nguyen *et al.* 2015), allowing the best-fitting substitution model to be selected automatically (Kalyaanamoorthy *et al.* 2017). Each tree was screened by PhyloTreePruner (Kocot *et al.* 2013) and any OGs containing paralogues were discarded. If two representative sequences were present for any species (ie., "in-paralogs") after this paralog screening step, the longest of the two sequences was retained and the other discarded. I realigned the protein sequences of each remaining OG using MAFFT.

Supermatrix approach

I trimmed spuriously aligned regions from each alignment using trimAl (Capella-Gutiérrez *et al.* 2009). The trimmed alignments were concatenated using catfasta2phyml (available from: https://github.com/nylander/catfasta2phyml) to form a supermatrix. I inferred the species tree using maximum likelihood (ML) using IQ-TREE, with the general-time reversible (GTR) substitution model with gamma-distributed rate variation among sites (+Γ) along with 1000 ultrafast bootstraps. Bayesian inference was carried out using the site-heterogeneous CAT-GTR+Γ

substitution model (Lartillot & Philippe 2004) with gamma-distributed rate variation among sites) implemented in PhyloBayes MPI (Lartillot *et al.* 2013), with four independent Monte Carlo Markov chains (MCMC). Convergence was assessed using Tracer (Rambaut *et al.* 2007). A posterior consensus tree was estimated using samples from both chains, with the initial 10% of all trees discarded as burn-in. The resulting species trees were visualized using the iTOL web server (Letunic & Bork 2016).

Supertree approach

I inferred a gene tree for each OG using IQ-TREE, allowing the best-fitting substitution model to be selected automatically. I provided the resulting gene trees to ASTRAL-III (Zhang *et al.* 2018) to estimate the species tree. Recent studies have suggested that the accuracy of species trees estimated by ASTRAL-III may be improved by collapsing nodes with low support in the input gene trees (Zhang *et al.* 2018). I created two new sets of gene trees by collapsing weakly supported nodes (bootstrap values below 10 and 20, respectively) into polytomies using newick utilities (Junier & Zdobnov 2010). Both resulting topologies were identical to the one inferred using uncollapsed genes trees. I used IQ-TREE and the GTR+Γ substitution model to estimate branch lengths in substitutions per site for the topology recovered by ASTRAL-III.

Assessing support for contentious relationships

gCF values for each branch in all recovered species trees were calculated using IQ-TREE and all 2,869 ML-estimated gene trees. I performed multiple linear regression analysis with gCF values as the independent variable and branch length and distance from root (both in substitutions per site) as dependant variables using R. I also performed constrained gene tree searches in IQ-TREE under four phylogenetic hypotheses (Fig. 2). For each gene, support for the resulting topologies was assessed using AU tests and the tree with the highest probability was recorded.

Chapter 5

The evolution of genome size and content in *Caenorhabditis*

Abstract

Eukaryotic genomes vary extensively in their size and content. Studies in several lineages suggest that variation in genome size is the result of a diverse range of molecular processes, including whole-genome duplication, changes in gene structure, and proliferation of repetitive elements, which are controlled by a range of evolutionary drivers, including reproductive mode, population size and neutral stochasticity. In *Caenorhabditis*, there has been a particular focus on understanding the genomic consequences of a switch in reproductive mode from obligate outcrossing to self-fertile hermaphroditism. Here, I use draft genome sequences of 48 *Caenorhabditis* species to investigate the evolution of genome size and content in the genus. I show that genome size varies extensively, from the 48 Mb genome of *C. drosophilae* to the 165 Mb genome of *C.* sp. 54. I show that changes in the number of protein-coding genes and proportion of repetitive DNA are significantly correlated with genome size. I reveal that variation in gene number is linked to the expansion and contraction of gene families and identify several gene families that co-vary with gene number. Interestingly, while I find that many *Caenorhabditis* species have undergone extensive intron loss during their evolution, there is no correlation between genome size and intron number. My results represent a substantial contribution to our understanding of genome evolution in this important genus of nematodes.

Introduction

Eukaryotic genomes vary extensively in their size, content, and structure. The genome of the nematode *Pratylenchus coffeae*, for example, spans 20 Mb and contains just 6,000 protein-coding genes (Burke *et al.* 2015), while the genome of the axolotl *Ambystoma mexicanum* is more than 1,500 times larger, spanning 32 Gb (Nowoshilow *et al.* 2018). The apparent lack of any correlation between genome size and organismal complexity or gene number has been termed the "C-value paradox" (Cavalier-smith 1985). In recent years, genomes for species across the tree of life have become available, revealing that variation in genome size and content results from a diverse range of processes, including whole-genome duplication (Cui *et al.* 2006), changes in gene content and structure (Yoshida *et al.* 2017), and proliferation of transposable elements (Naville *et al.* 2019). However, the nature of the evolutionary forces that drive these changes remain obscure. It has been argued that changes in genome size are non-adaptive in nature and ultimately the product of neutral population genetic processes.(Petrov 2001; Lynch & Conery 2003). Briefly, this argument posits that, in species with small effective population sizes, the process of neutral genetic drift leads to the accumulation of genomic features that would otherwise be removed from the genome by purifying selection (Lynch & Conery 2003). This appears to provide an explanation for why the genomes of species with low effective population sizes, such as mammals, tend to be relatively large, contain larger and more numerous introns, and contain more transposable elements than those with larger effective population sizes (Lynch & Conery 2003)

In *Caenorhabditis*, there has been a particular interest in the genomic consequences of switch in reproductive mode. Self-fertile hermaphrodites have independently evolved from obligately outcrossing gonochoristic ancestors three times in the genus (Kiontke *et al.* 2011), including in *C. elegans*. This change in reproductive mode is expected to lead a reduction in genome size as genomic features associated with mating are lost (Thomas et al. 2012). In contrast, the evolution of hermaphroditism is expected to lead to a reduction in the effective population size which will, in turn,

lead to the proliferation of mildly deleterious elements, such as transposons, leading to an increase in genome size (Charlesworth & Wright 2001; Fierst et al. 2015). Previous studies have revealed that genomes and transcriptomes of hermaphroditic species are smaller than those of their outcrossing relatives (Thomas *et al.* 2012; Fierst *et al.* 2015). A detailed comparison of the closely related sister taxa *C. nigoni* and *C. briggsae* found that the genome of *C. briggsae* has undergone extensive contraction since the evolution of hermaphroditism, largely due to loss of genes with male-biased expression (Yin *et al.* 2018). However, in the other two self-fertile species, *C. elegans* and *C. tropicalis*, significant loss of genes involved in mating does not appear to explain the difference in genome size between their outcrossing sisters, *C. inopinata* and *C. wallacei (Kanzaki et al. 2018)*; Erich Schwarz *pers. comm.*). In addition, some outcrossing species have genomes that are significantly smaller than those of the three hermaphroditic species (Stevens *et al.* 2019), suggesting that factors other than reproductive mode play roles in driving changes in genome size in *Caenorhabditis*.

Here, I use genome sequences of 48 *Caenorhabditis* species to investigate the evolution of genome size and content in the genus. I find evidence for an over three-fold variation in genome size in the genus, ranging from 48 Mb (*C. drosophilae*) to 165 Mb (*C.* sp. 54). I show that changes in both the number of protein-coding genes and proportion of repetitive DNA are highly correlated with genome size. I reveal that changes in gene family size underlie variation in gene number and identify several large gene families whose size covaries with gene number. Interestingly, I find no correlation between genome size and intron number. I reveal that many *Caenorhabditis* species, including *C. elegans*, have undergone extensive intron loss during their evolution. My results represent a substantial contribution to our understanding of genome evolution in this important genus of nematodes.

Results

Extensive variation in genome size in *Caenorhabditis*

To avoid being misled by artefacts of genome assembly, I first identified and excluded assemblies which retained uncollapsed duplication, as evidenced by their containing an excess of loci that were duplicated in the genome but single-copy in most other *Caenorhabditis* genomes. These likely had artificially inflated assembly spans and protein-coding gene counts. The genomes of ten species were identified as containing a higher-than-average number of duplicated loci and were excluded (Fig. S1). Six of these were identified previously (see chapter 1). My final dataset contained genomes for 48 *Caenorhabditis* species, all of which contained low levels of duplication.

Genome size is highly variable within the genus, ranging from the 48 Mb genome of *C. drosophilae* to the 165 Mb genome of *C.* sp. 54 (Fig. 1A). The pattern of variation in genome size shows a strong phylogenetic signal (Pagel's lambda=0.964, P<0.001; Blomberg's K=0.3254, P=0.006), with closely related taxa possessing genomes of similar size. For example, the sister taxa *C. drosophilae* and *C.* sp. 2 have genome sizes of 48 and 50 Mb respectively, while the sister taxa *C. portoensis* and *C.* sp. 27 have a genome sizes of 161 and 160 Mb, respectively. Different clades have different average genome sizes: the *Angaria* group has an average genome size of 87 Mb (± 15 Mb), while the *Elegans* group has an average genome size of 111 Mb (± 21 Mb). Consistent with previous analyses (Kanzaki *et al.* 2018; Yin *et al.* 2018), the genomes of all hermaphroditic species (*C. elegans*, *C. briggsae,* and *C. tropicalis*) are smaller than the genomes of their outcrossing sister taxa (*C. inopinata*, *C. nigoni*, *C. wallacei*, respectively). However, the differences between each species pair (1.3 Mb - 22.7 Mb) are small in comparison to the variation present in the genus as a whole.

To identify which genomic features underlie this variation, I used phylogenetic generalised least squares analysis (PGLS) to test for correlation between genomic

features and genome size while controlling for phylogenetic relatedness. I find that protein-coding gene number and estimated repeat content are positively and significantly correlated with genome size (r^2=0.51; $P < 0.001$ and r^2=0.73; $P < 0.001$, respectively; Fig. 1B,C). The smallest genomes in the genus, those of *C. drosophilae* and *C.* sp. 2, possess some of the smallest protein-coding gene sets, with 13,712 and 13,557, respectively, and between 3.5-4.6 Mb of repetitive sequence. In contrast, the largest genome, that of *C.* sp. 54, contains nearly three times as many predicted genes (35,881) and over 41 Mb of repetitive sequence. Interestingly, I find that while intron size is significantly correlated with genome size ($r^2 = 0.24$; $P = 0.04$), the number of introns per gene shows no correlation with genome size ($r^2 = 0.04$; $P = 0.19$; Fig. 1D,E).

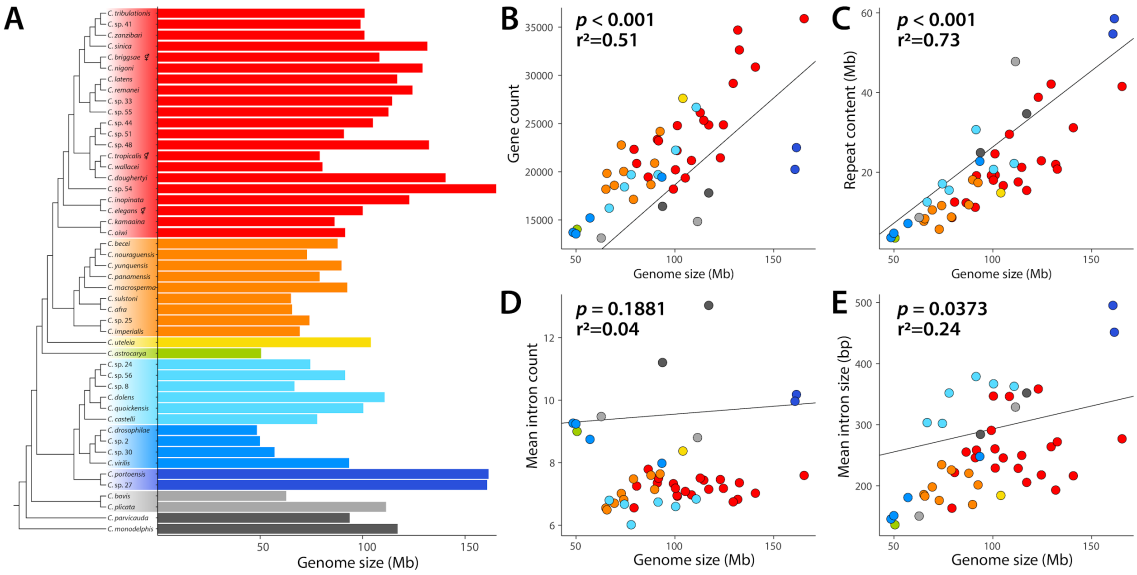

Figure 1: Extensive variation in genome size in *Caenorhabditis*

A: Genome size is highly variable in 48 *Caenorhabditis* species and shows strong phylogenetic signal (Pagel's lambda=0.964, *P*<0.001; Bloomberg's K=0.3254, *P*=0.006). **B-E**: PGLS analysis of protein-coding gene count, estimated repeat content, mean intron number per gene, and mean intron size versus genome size. Clades are highlighted. To limit the effect of gene prediction artefacts, I calculated mean intron count and size in a set of 866 single-copy orthologues present in all 48 species.

Evolution of gene content in *Caenorhabditis*

My finding that protein-coding gene number is correlated with genome size suggests that multiple lineages have undergone gene gain and loss during their evolutionary histories. Using the orthology clustering set described previously, I determined the proportion of multi-copy (genes which clustered alongside others from the same species and with at least one gene from another species), single-copy (genes which did not cluster with any others from the same gene species but with at least one gene from another species), and unique (genes that did not cluster with genes from any other species) in each gene set. I find that, as gene number increases, the proportion of multi-copy genes also increases (Fig. 2A; Fig. S2A,B). For example, 58% of genes (20,920) in the largest gene set (*C.* sp. 54) are multi-copy, while only 22% of genes (2,897) are multi-copy in the smallest gene set (*C. bovis*). The proportion of species-specific genes (unique genes) was also significantly correlated with protein-coding gene number (Fig. 2A; Fig. S2C). However, species-specific genes constitute of relatively minor faction of each gene set (mean of 10% ±6%).. This suggests that variation in protein-coding gene number in *Caenorhabditis* is primarily the result of expansion and contraction of existing gene families, rather than *via* the gain and loss of large numbers of novel genes.

I investigated two gene families that are known notably expanded in *C. elegans* relative to other lineages (Robertson and Thomas 2006; Antebi 2006) to assess whether the size of these families was correlated with number of protein-coding genes in each species. G protein-coupled receptors (GPCRs) are a large family of transmembrane proteins many of which perform chemosensory roles in *C. elegans* (Robertson 1998). Nuclear hormone receptors (NHRs) are a large family of transcription factors with diverse roles in nematode metabolism, development and homeostasis (Taubert *et al.* 2011). GPCRs and NHRs constitute a substantial fraction of the *C. elegans* gene set (7% and 3%, respectively). I compared the number of these in the gene sets of all 48 species using PGLS analysis. The number of GPCRs and NHRs in each species is both positively and significantly correlated with the number of

protein coding genes (r^2=0.23; P < 0.001 and r^2=0.35; P < 0.001, respectively; Fig. 2B,C). In addition, I identified several other gene families whose size is positively correlated with gene count (Fig. S3). The family containing the *C. elegans* worm-specific Argonaute (WAGO) proteins *hrde-1, nrde-1,wago-10* and *wago-11* was significantly correlated with gene number (r^2=0.51; P < 0.001). WAGOs are part of the endogenous RNAi pathway in *C. elegans* which plays a key role in directing chromatin modulation and thus in repressing expression of genes, including transposons (Billi *et al.* 2014). Further analysis revealed that the total number of WAGOs in each species was highly correlated with protein-coding gene number (r^2=0.51; P < 0.001; Fig. 2C)

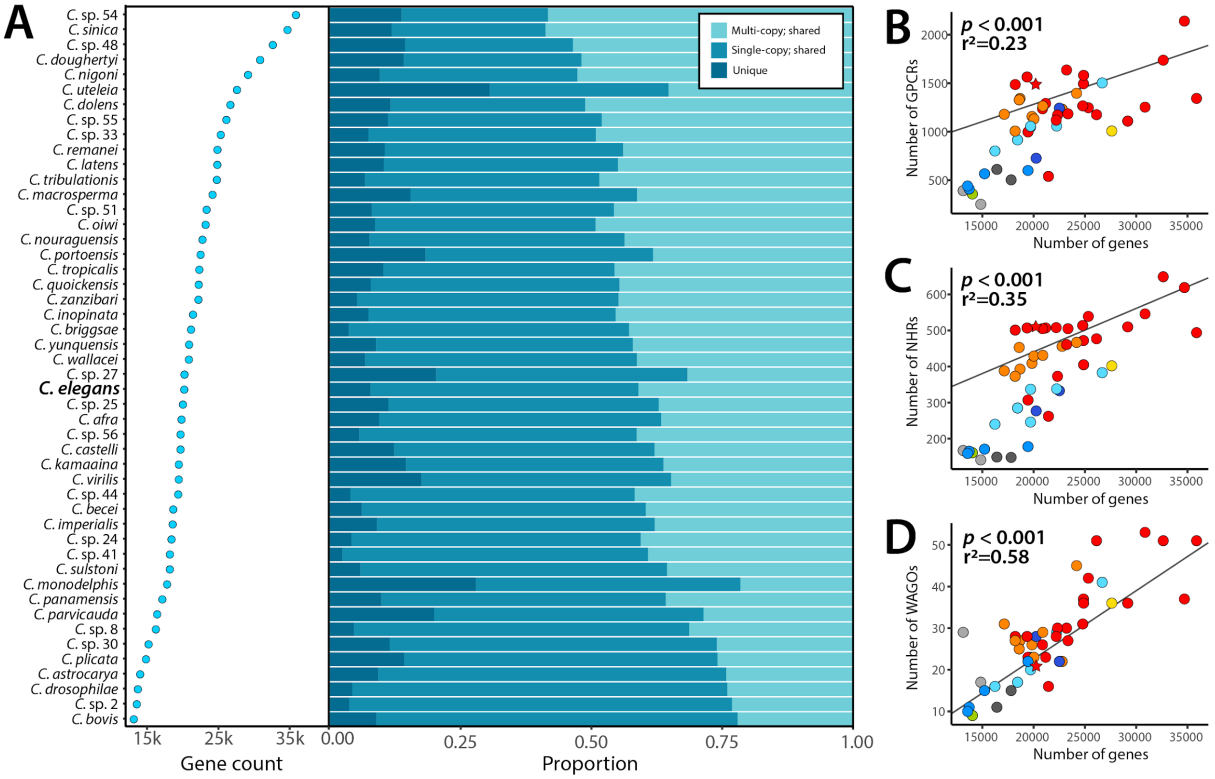

Figure 2: Protein-coding gene content evolution in *Caenorhabditis*

A: Protein-coding gene count and the proportion of species-specific genes and multi-copy and single-copy genes that are shared with at least one other species. Species are ordered by gene count.

B-E: PGLS analysis of the number of GPCRs, NHRs, and WAGOs versus genome size. Clades are highlighted as per the phylogeny in Fig. 1A. Stars indicate values for *C. elegans*.

Extensive intron loss in *Caenorhabditis*

I sought to further investigate variation in intron abundance in the genus. Given that the high levels of duplication in the genomes excluded previously would not interfere with analyses of intron gain and loss, I opted to include all 58 species including the outgroup *Diploscapter* species. I defined 659 single-copy orthologues present in all 58 species and counted the number of introns in each gene. I find that intron abundance is highly variable in the genus, with the genomes of basal and outgroup taxa containing substantially more introns than the genomes of ingroup taxa (Fig. 3A). The 659 loci in the most early diverging species, *C. monodelphis,* contain an average of 12.4 introns per gene (a total of 8,192 introns), while their orthologues in *C. elegans* contain an average of 6.39 introns per gene (a total of 4,212 introns). The pattern shows a strong phylogenetic signal (Pagel's lambda=0.994, $P<0.001$; Blomberg's K=8.205, $P=0.001$), with different clades containing different abundances (Fig. 3A). For example, the *Portoensis* group species contain an average of 9.45 introns per gene, while *Angaria* group species contain an average of 5.9 introns per gene.

I sought to determine whether this pattern was the product of extensive intron gain in some groups or extensive intron loss in others. Using nucleotide alignments of all 659 single-copy orthologues, I identified 13,812 intron-containing sites at which at least 50 species had aligned sequence. I identified orthologous introns as those existing at the same site, and used Dollo parsimony to infer the number of putative gain and loss events on the *Caenorhabditis* phylogeny. I infer that a minimum of 6,708 introns existed in the last common ancestor of all *Caenorhabditis* species and, therefore, that many species have undergone extensive intron loss during their evolution (Fig. 3A). A substantial proportion of the remaining intron sites (4,665) appear to be species-specific gains. However, the inferred rate of intron gain in these terminal branches is substantially higher than in internal branches (mean rates of 140.6 and 81.9 gains per 0.1 substitutions per site, respectively; Fig. 3C), suggesting many of these specific-specific gains are artefacts due to misalignment or gene prediction errors. By considering the internal branches only, I estimate that

approximately 11 intron losses have occurred for every intron gain (14,398 losses and 1,321 gains) since the last common ancestor of all *Caenorhabditis* species.

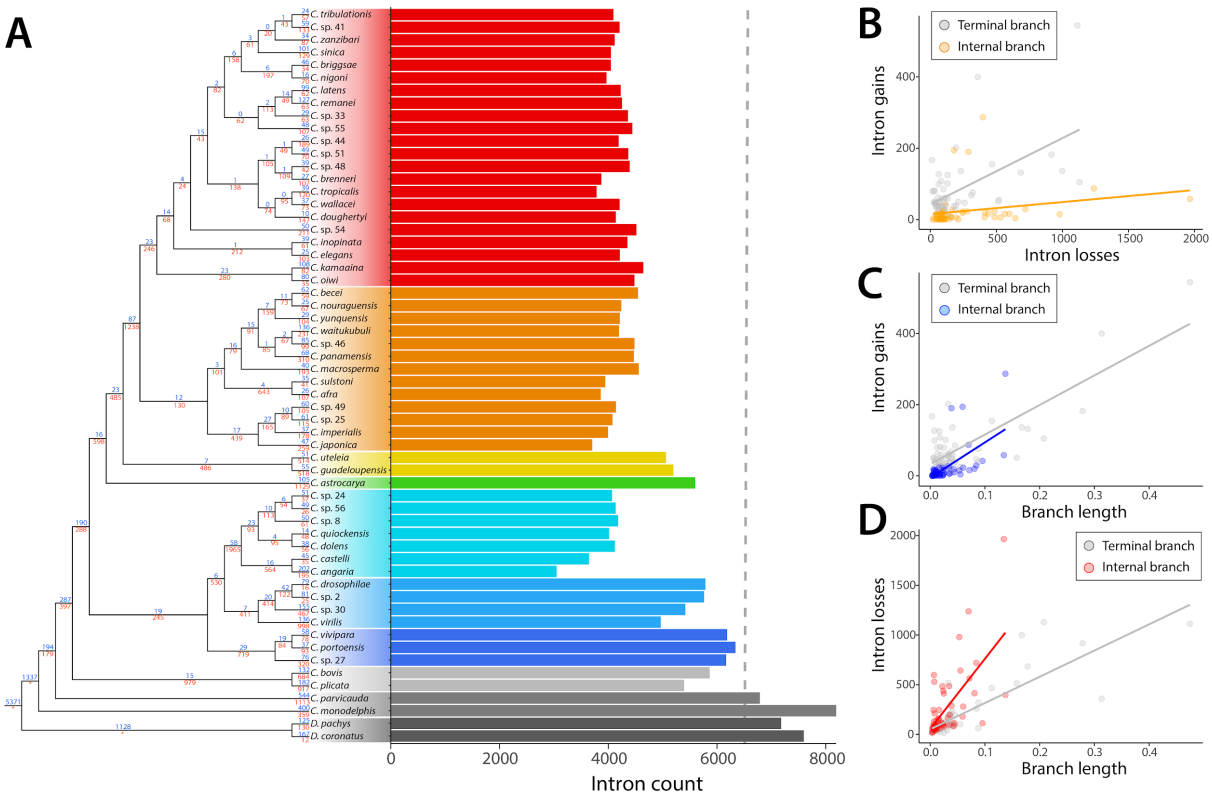

Figure 3: Extensive intron loss in *Caenorhabditis*

A: Intron abundances across the phylogeny. Bars represent the number of introns in 659 single-copy orthologues present in all 58 species. Branch annotations show number of inferred gains (blue) and losses (red). Branch lengths are not shown. * represents branches were it loss counts were not inferred. Intron sites at which fewer than 50 species had aligned sequence were excluded. Dashed line represents inferred number of ancestral *Caenorhabditis* introns (6,508). **B:** Number of gains and losses each branch. Linear regression lines are shown for internal and terminal branches separately. **C:** The number of intron gains on internal branches is significantly correlated with branch lengths ($r^2 = 0.353$; $P < 0.01$). Branch lengths are in amino acid substitutions per site. Linear regression lines are shown for internal and terminal branches separately. **D:** The number of intron losses on internal branches is significantly correlated with branch length ($r^2 = 0.412$; $P < 0.01$). Branch lengths are in amino acid substitutions per site. Linear regression lines are shown for internal and terminal branches separately.

Discussion

By comparing species from across the genus *Caenorhabditis, I* found that genome size is highly variable and is highly correlated with protein-coding gene number and repeat content. I also found that protein-coding gene number is highly correlated with the proportion of multi-copy genes, and identified several gene families whose size is positively correlated with protein-coding gene number. I investigated the origin of variation in intron abundance in the genomes of these species, and reveal that many species have undergone extensive intron loss during their evolution.

My finding that genome size and repeat content are highly correlated is consistent with studies in many different eukaryotic lineages, including mammals, *Drosophila*, and plants (Drosophila 12 Genomes Consortium *et al.* 2007; Tenaillon *et al.* 2011; Platt *et al.* 2018). In contrast, based on results from vertebrates and plants, it has been argued that changes in gene number play a relatively minor role in genome size changes in eukaryotes (Kidwell 2002). My finding that protein-coding gene number, and changes in the number of GPCRs, NHRs and WAGOs specifically, are correlated with genome size suggest that this is not true for *Caenorhabditis*. The GPCR expansion in *C. elegans* has been noted to imply a surprisingly rich chemosensory capacity in this species. The discovery that GPCRs are similarly expanded in other *Caenorhabditis* genomes suggests that selection for increased complexity of this sensory modality may be a common driver of both GPCR family size and, incidentally, genome size. The expanded NHR family of ligand-binding transcription factors in *C. elegans* may be related to the evolution of responses to the complex ascaroside family of secreted signalling molecules. Thus increased NHR diversity in other *Caenorhabditis* species may also be related to selection for more complex ascaroside communication and sensing, and again have resulted in genome expansion. While GPCRs and NHRs are both notably expanded in *C. elegans*, I was surprised to find that the number of WAGOs was correlated with gene number. It is possible that, as larger genomes contain more transposons, the increase in the number of WAGOs is related to an increased requirement for transposon silencing, and thus may be a

pattern consequent to genome expansion caused by proliferation of mobile elements. It remains unclear why changes in gene content generally play such a major role in genome size changes in *Caenorhabditis*. It is possible that, because *Caenorhabditis* genomes are relatively compact and gene-dense, any duplication or deletion event is more likely to involve a genic region than it would in larger, less gene-dense genomes.

While I did not explicitly address the relationship between genome size and reproductive mode here, the genomes of all three hermaphroditic species are smaller than their outcrossing sister taxa. For *C. nigoni and C. briggsae*, a reduction of 21 Mb in the genome of the hermaphroditic species *C. briggsae* appears to be directly related to reproductive mode, with a substantial loss of genes which show a male-biased expression (Yin *et al.* 2018). In contrast, the 23 Mb size difference between *C. elegans* and *C. inopinata* does not appear to be the result of extensive gene loss in *C. elegans (Kanzaki et al. 2018)*. Rather, it appears that the *C. inopinata* genome has expanded due to the proliferation of various transposable elements (Kanzaki *et al.* 2018). In addition, the genome of hermaphroditic species *C. tropicalis* is only 1.3 Mb smaller the outcrossing sister taxon, *C. wallacei*, and *C. tropicalis* is predicted to have more protein-coding genes than *C. wallacei*. A detailed analysis on the difference between the genomes of these two species is currently underway (Erich Schwarz *pers. comm.*). Given these and my own findings, genome size variation within *Caenorhabditis* appears to be driven by multiple interacting mechanisms. The relative contributions of contrasting molecular processes, such as gene loss versus gene family expansion, or intergenic contraction versus repeat expansion, and their evolutionary drivers, such as a switch in reproductive mode, must differ between different lineages in the genus.

My finding that intron losses outnumber intron gains is consistent with previous studies in *Caenorhabditis* which have used individual or small numbers of genes (Robertson 1998; Cho *et al.* 2004; Kiontke *et al.* 2011). A predominance of intron losses over gains has also been reported in several other eukaryotic lineages,

including mammals, plants, and *Drosophila (Roy et al. 2003; Coulombe-Huntington & Majewski 2007; Roy & Penny 2007)*. However, one potential limitation of my analysis is in the use of Dollo parsimony. In the context of intron gain and loss, Dollo parsimony assumes that independent insertions of introns into the same site do not occur. While this assumption is likely to be valid for the vast majority of intron sites, it is possible that a small proportion of introns I define as being orthologous are instead the product of multiple, independent gains in the same position. It is also likely that Dollo parsimony will have resulted in an underestimation of the number of ancestral introns, as any ancestral intron that has been lost independently from *C. monodelphis* and the *Diploscapter* species but retained by two or more of the other lineages will be considered to have been gained elsewhere. Therefore, in addition to parsimony-based approaches, I plan to include probabilistic approaches of inferring gain and loss events in future analyses.

The mechanisms of intron gain and loss in eukaryotes remain obscure (Fedorov *et al.* 2003). Although several models of intron loss have been proposed, including recombination of the intron-containing locus with reverse-transcribed spliced mRNA (Sverdlov *et al.* 2004), clear empirical evidence remains lacking. Mechanisms of intron gain have proved even more difficult to elucidate (Yenerall & Zhou 2012). In *Caenorhabditis*, many reported instances of intron gain were subsequently found to be the result of multiple independent losses (Coghlan & Wolfe 2004; Roy & Penny 2006). I did not investigate potential mechanisms in this analysis, but the high rate of apparent gains in the terminal branches suggests that any attempt to infer mechanism would first require the elimination of artefacts arising from misalignment and gene prediction error. I did, however, find that intron abundance was not correlated with genome size. This suggests that the evolutionary forces that govern rates of intron gain and loss are distinct from those that are responsible for genome expansion and contraction or that they act over different timescales. This conflicts with theoretical predictions which suggest that intron abundance, like genome size, is fundamentally determined by effective population size (Lynch 2002).

Many questions therefore remain about this particular feature of genome evolution in *Caenorhabditis* and in eukaryotes generally.

Methods

Orthology clustering and identification of duplicated genomes

Details of all data used in the orthology clustering analysis are available in Table S1. I collected the protein sequences predicted from the genomes of 56 *Caenorhabditis* species and two outgroup taxa (*Diploscapter coronatus* and *Diploscapter pachys*) and selected the longest isoform of each gene. OrthoFinder (Emms & Kelly 2015) was used to cluster all protein sequences into putatively orthologous groups (OGs). To identify genomes which were highly duplicated, I selected OGs which were present in all species and had an average count of <1.2. I calculated duplication scores by dividing the number of loci for each species by the number of OGs, and used the scores to identify species that contained higher-than-average levels of duplication. Ten species (*C. angaria, C. brenneri, C. guadeloupensis, C. japonica, C.* sp. 46, *C.* sp. 49, *C. vivipara, C. waitukubuli, Diploscapter coronatus,* and *Diploscapter pachys*) had duplication scores > 1.15 and were excluded. I reclustered the protein sequences of the remaining species into OGs using OrthoFinder.

Gene family analysis

I classified each gene in each species as multi-copy (genes which clustered alongside at least one other gene from the same gene set and with at least one gene from another species), single-copy (genes which did not cluster with any other genes from the same gene set but with at least one gene from another species) or unique (those that did not cluster with genes from any other species) and compared their proportions using PGLS. To identify gene families whose size was correlated with protein-coding gene number, I calculated counts for each OG using KinFin (Laetsch & Blaxter 2017b) and compared them with protein-coding gene number using PGLS. GPCRS, NHRs and WAGOs were identified using InterProScan. Briefly, I searched the longest isoform of each protein-coding gene in all 35 species against the Pfam and

SignalP databases using InterProScan (Jones *et al.* 2014) and provided the resulting annotations to KinFin.

Phylogenetic comparative methods

I used RepeatModeller (Smit & Hubley 2010) to identify repetitive sequences in each genome independently. The resulting repeat libraries were provided to RepeatMasker (Smit *et al.* 1996) to estimate the span of repetitive DNA in each species. To limit the effect of gene prediction artefacts, I calculated mean intron count and size in a set of single-copy orthologues present in all 48 species. I conducted PGLS analysis of genome size and content using the R packages ape (Paradis & Schliep 2018), caper (Orme *et al.* 2013), and phytools (Revell 2012) using the Brownian model of evolution and species tree. Statistical tests for phylogenetic signal (Pagel's lambda and Blomberg's K) were conducted using the R packages phytools, ape, geiger (Pennell *et al.* 2014), and nlme (Pinheiro *et al.* 2012).

Intron gain and loss

I opted to include all species in our analyses of intron gain and loss as the outgroup *Diploscapter* species are particularly important for inferring ancestral intron counts. Using the orthology set containing 58 species, described previously, I selected OGs which were, on average, single-copy and present in at least 75% of species using KinFin (Laetsch & Blaxter 2017b). To identify paralogous sequences, I aligned the protein sequences of each selected OG using MAFFT (Katoh & Standley 2013) and generated a maximum likelihood tree along with 1000 ultrafast bootstraps (Hoang *et al.* 2018) using IQ-TREE (Nguyen *et al.* 2015), allowing the best-fitting substitution model to be selected automatically (Kalyaanamoorthy *et al.* 2017). Each tree was screened by PhyloTreePruner (Kocot *et al.* 2013) and any OGs containing paralogues were discarded. If two representative sequences were present for any species (ie., "in-paralogs") after this paralog screening step, the longest of the two sequences was retained and the other discarded. I realigned the protein sequences of each remaining OG using MAFFT. The protein alignments were translated into nucleotide alignments

using PAL2NAL (Suyama *et al.* 2006). I collected intron positions from the GFF3 files of each species using customs scripts. I identified intron sites in each alignment where at least 50 species had aligned sequence, and declared orthology between introns if they existed in the same position in the alignment. I used a custom Python script (with extensive use of the ETE3 module (Huerta-Cepas *et al.* 2016a)) to infer putative gain and loss events on the phylogeny using Dollo parsimony.

Chapter 6

General Discussion

Thesis Overview

The central aim of this thesis was to generate a high-quality comparative genomic dataset for *Caenorhabditis* and to subsequently investigate the patterns and processes of genome evolution in the genus. Ultimately, I sought to create a resource that could be used to place *C. elegans* and the vast body of associated research within an evolutionary context.

In chapter 2, I presented draft genome sequences for 38 *Caenorhabditis* species. Using a range of quality metrics, I demonstrated that a majority of these data are of sufficient quality and completeness for use in downstream analyses of genome evolution. I also demonstrated the utility of long-read sequencing data for generating genome assemblies of high contiguity.

In chapter 3, I presented the draft genome of *C. bovis*, an unusual and understudied *Caenorhabditis* species which appears to live parasitically in the ears of cattle in Eastern Africa. With the help of local veterinary practitioners and scientists, I reisolated C. bovis from an infected adult Zebu in Western Kenya and used a portable sequencing platform to sequence the *C. bovis* genome in a nearby field laboratory. I revealed several features of the *C. bovis* genome that may play a role in its unusual lifestyle.

In chapter 4, I conducted the most comprehensive reconstruction of the *Caenorhabditis* phylogeny to date. I found that the majority of relationships are

recovered with high support regardless of method and performed hypothesis testing approaches to further investigate contentious regions of the phylogeny. The results of this chapter provide a phylogenetic framework that will be fundamental to future evolutionary studies in *Caenorhabditis*.

Finally, in chapter 5, I investigated the evolution of genome size and content within the context of the *Caenorhabditis* phylogeny. I reveal that genome size is highly variable in the genus, and is largely the product of changes in protein-coding gene content and proportion repetitive DNA. I also revealed that many *Caenorhabditis* species have undergone extensive intron loss during their evolution.

In this chapter, I will place this work within the context of previous work on genome evolution in other eukaryotic genera. I will also discuss the importance of this dataset for the *C. elegans* research community. Lastly, I highlight several remaining questions about genome evolution in *Caenorhabditis* and in eukaryotes generally.

Comparative genomics of other eukaryotic genera

With genomes for 59 of the 64 known species now sequenced (chapter 1; Taisei Kikuchi *pers. comm.*), the genus *Caenorhabditis* now has one of the richest comparative genomics datasets of any eukaryotic genus.

The fruitfly *Drosophila melanogaster* is an important model for animal genetics with a long history of use in biological research. *D. melanogaster* is one of over 1,500 species of *Drosophila* that have a global distribution (Drosophila 12 Genomes Consortium *et al.* 2007). Like the *C. elegans* genome, the *D. melanogaster* genome has been extensively annotated by a large community of researchers. In 2004, the genomes of 12 *Drosophila* species were published and used to study the evolution of genes and chromosomes within the context of the *Drosophila* phylogeny (Drosophila 12 Genomes Consortium *et al.* 2007). Similar to *Caenorhabditis*, these analyses revealed genome size in *Drosophila* is highly variable, ranging from 130 Mb to 364 Mb. The majority of the difference in genome size between the species is due to differences in the number of transposable elements. In contrast to *Caenorhabditis*, the variation in protein-coding gene number between these species was relatively small (13,000 - 17,000) (Drosophila 12 Genomes Consortium *et al.* 2007). These genomes sequences have since become fundamental to the *Drosophila* research community and have been used to study the evolution of several features of genome biology, including miRNAs (Stark *et al.* 2007), transposable elements (Yang & Barbash 2008), and cell signalling pathways (Alvarez-Ponce *et al.* 2009). Recently, highly-contiguous assemblies of 15 *Drosophila* species, 14 of which had been sequenced previously, were generated using long-read sequencing (Miller *et al.* 2018).

Other genera for which large comparative genome datasets have been assembled include the yeast genus *Saccharomyces* and the butterfly genus *Heliconius*. In

Saccharomyces, as in *Caenorhabditis*, high levels of incongruence exist among gene trees and support topologies that are in conflict with most likely species tree (Rokas *et al.* 2003). Highly-conserved regions of these genomes have been used identify previously-unidentified genes and regulatory elements in *Saccharomyces cerevisiae* genome (Kellis *et al.* 2003). Recently, the genomes of 45 *Heliconius* species were sequenced (Kozak *et al.* 2018). *Heliconius* species frequently form hybrids in nature and their genomes show evidence of extensive gene-flow between related species, including adaptive introgression of loci involved in wing patterning (Martin *et al.* 2013; Kozak *et al.* 2018). As a result, a large proportion of gene trees in this genus support evolutionary histories that are distinct from the species tree (Kozak *et al.* 2018).

The *Caenorhabditis* dataset described here thus adds to the base of knowledge that can be used to explore shared patterns of genome evolution in animals. While *Caenorhabditis* nematodes are ecdysozoans, like the insects *Heliconius* and *Drosophila*, the addition of this comparator significantly broadens the diversity of available animal datasets. Importantly, as most if not all *Caenorhabditis* species share many of the same experimenter-friendly features as *C. elegans*, and are likely to be amenable to many if not all of the experimental manipulation tookits developed for *C. elegans* (e.g. RNAi, CRISPR gene editing), the genus is particularly well placed for reverse and forward genetic analysis of traits of interest.

A resource for the *C. elegans* research community

While our findings represent a significant contribution to the understanding of genome evolution in *Caenorhabditis*, the most significant outcome of this work is a resource that we have created for the *C. elegans* research community. For the majority of its time as a model organism, *C. elegans* has been studied in isolation and the associated research community has lacked the resources to place *C. elegans* into an evolutionary context. Thanks to efforts by our collaborators, recent years have seen significant progress in our understanding of *Caenorhabditis* diversity. By sequencing the genomes of all newly discovered species, we have made a significant contribution to understanding the biology of these new species far more easily accessible. With these new species and their genomic resources in hand, it is now possible to study any *C. elegans* system across 60 closely related species.

To facilitate the use of our data, I have been proactive in releasing our genome sequences ahead of publication *via* our project website (caenorhabditis.org), enabling any researcher worldwide to browse, BLAST and download our data. Recently, I have also made orthology sets and gene trees available. As a result, our data are already being used by the *C. elegans* research community and have facillitated studies of the evolution of asexual reprorduction (Lamelza & Ailion 2016), the evolution of developmental pathways (Barkoulas *et al.* 2016), piRNA biogenesis (Beltran *et al.* 2019), and RNA interference (Braukmann *et al.* 2019). All of our raw genomic data will be submitted to the relevant INSDC databases and draft genomes will eventually be migrated to their permanent home in WormBase (Lee *et al.* 2018) and WormBase parasite (Howe *et al.* 2017). We also intend to publish all genome sequences described here so that they are free to be used by the community. I hope that these efforts mean that our data will add to the already expansive arsenal of tools available to the *C. elegans* research community to understand the biology of this important model organism.

Remaining questions surrounding genome evolution in the genus *Caenorhabditis*

Many interesting questions surrounding genome evolution in Caenorhabditis concern the evolution of large-scale genome organisation. The five C. elegans autosomes show a clear structure with distinct domains, termed arms and centers, which differ in their composition and recombination rate (C. elegans Sequencing Consortium 1998; Rockman & Kruglyak 2009). This pattern of genome organisation appears to be conserved in other species, with similar organisations reported for C. briggsae and C. nigoni (Stein et al. 2003; Hillier et al. 2007; Yin et al. 2018). In addition, despite the extremely high rates of intrachromosomal rearrangement, chromosomal linkage groups are highly conserved in Caenorhabditis and rhabditine nematodes generally (Stein et al. 2003; Tandonnet et al. 2019). Furthermore, the karyotype of n=6 also appears to be conserved in most rhabditine nematodes (Walton 1959). Investigating the origins of these patterns of genome organization in Caenorhabditis would require chromosome-scale draft genomes for many species in the genus. The majority of draft genomes presented in this thesis were sequenced using short-read technology only. As a result, most are highly fragmented and large-scale patterns of genome organisation are completely obscured. I have shown that long-read sequencing technology can substantially improve assembly contiguity and, in the case of C. bovis, lead to the assembly of complete chromosomes. Analysis of these complete C. bovis chromosomes revealed a distribution of transposable elements that was markedly different from C. elegans. Chromosome-scale draft genomes for further species in the genus would shed light on which processes govern the evolution of these patterns. Consequently, I propose resequencing all species using long-read sequencing technology.

While I presented high-quality protein-coding gene predictions for all genomes presented here, there are many other genomic features that are relevant to the biology of these organisms. Non-coding RNAs (ncRNAs) are known to play diverse

roles in the biology of many organisms, including *C. elegans (Stricklin et al. 2005)*. While small RNA sequencing datasets have used to identify ncRNA loci in several *Caenorhabditis* species (Sarkies *et al.* 2015; Beltran *et al.* 2019), we know very little about the location and nature of the ncRNA genes in the genomes of most species. In addition to ncRNAs, transposable elements are also key features of eukaryotic genomes. While I estimated the proportion of repetitive DNA in each genome (the majority of which is made up of transposable elements), I did not classify the resulting sequences into transposable element families. This was, in part, due to the fact that automated identification of transposable elements remains challenging. The majority of methods rely on homology to previously-classified sequences (such as those from *C. elegans*) (Tarailo-Graovac & Chen 2009), and such approaches become increasingly ineffective as phylogenetic distances increase. In future, I plan to perform a more comprehensive annotation of transposable elements in the of these species and explore how they have evolved and proliferated over time.

Concluding remarks

This thesis presents the results of a genus-wide genome sequencing project of the genus *Caenorhabditis*. I have employed the latest sequencing technology and bioinformatic approaches to generate draft genomes for 38 *Caenorhabditis* species and exploited these sequences to explore genome evolution in the genus. The data presented here are already being used by the *C. elegans* research community to perform detailed analyses of the evolution of specific biological systems. However, many questions surrounding the evolutionary origins of *C. elegans* still remain. In the coming years, I hope that our data are exploited by the *C. elegans* research community to help place this important nematode and the vast body of associated research within a rich evolutionary context. I am excited to hear what they find.

References

Albertson, D.G. & Thomson, J.N. (1982). The kinetochores of Caenorhabditis elegans. *Chromosoma*, 86, 409–428.

Alvarez-Ponce, D., Aguadé, M. & Rozas, J. (2009). Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 Drosophila genomes. *Genome Res.*, 19, 234–242.

Andersen, E.C., Gerke, J.P., Shapiro, J.A., Crissman, J.R., Ghosh, R., Bloom, J.S., *et al.* (2012). Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity. *Nat. Genet.*, 44, 285–290.

Andrews, S. & Others. (2010). FastQC: a quality control tool for high throughput sequence data.

Arakawa, K. (2016). No evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proc. Natl. Acad. Sci. U. S. A.*

Arcila, D., Ortí, G., Vari, R., Armbruster, J.W., Stiassny, M.L.J., Ko, K.D., *et al.* (2017). Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. *Nat Ecol Evol*, 1, 20.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., *et al.* (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, 19, 455–477.

Barkoulas, M., Vargas Velazquez, A.M., Peluffo, A.E. & Félix, M.-A. (2016). Evolution of New cis-Regulatory Motifs Required for Cell-Specific Gene Expression in Caenorhabditis. *PLoS Genet.*, 12, e1006278.

Barrell, B.G. & Sanger, F. (1969). The sequence of phenylalanine tRNA from E. coli. *FEBS Lett.*, 3, 275–278.

Barrière, A., Yang, S.-P., Pekarek, E., Thomas, C.G., Haag, E.S. & Ruvinsky, I. (2009). Detecting heterozygosity in shotgun genome assemblies: Lessons from obligately outcrossing nematodes. *Genome Res.*, 19, 470–480.

Bartley, D.J., McAllister, H., Bartley, Y., Dupuy, J., Ménez, C., Alvinerie, M., *et al.* (2009). P-glycoprotein interfering agents potentiate ivermectin susceptibility in ivermectin sensitive and resistant isolates of Teladorsagia circumcincta and Haemonchus contortus. *Parasitology*, 136, 1081–1088.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., *et al.* (2004). The Pfam protein families database. *Nucleic Acids Res.*, 32, D138–41.

Beltran, T., Barroso, C., Birkle, T.Y., Stevens, L., Schwartz, H.T., Sternberg, P.W., *et al.* (2019). Comparative Epigenomics Reveals that RNA Polymerase II Pausing and Chromatin Domain Organization Control Nematode piRNA Biogenesis. *Dev. Cell*, 48, 793–810.e6.

Bemm, F., Weiß, C.L., Schultz, J. & Förster, F. (2016). Genome of a tardigrade: Horizontal gene transfer or bacterial contamination? *Proc. Natl. Acad. Sci. U. S. A.*

Berriman, M., Coghlan, A. & Tsai, I.J. (2018). Creation of a comprehensive repeat library for a newly sequenced parasitic worm genome.

Billi, A.C., Fischer, S.E.J. & Kim, J.K. (2014). Endogenous RNAi pathways in C. elegans. *WormBook*, 1–49.

Blaxter, M. & Koutsovoulos, G. (2015). The evolution of parasitism in Nematoda.

Parasitology, 142 Suppl 1, S26–39.
Blaxter, M.L., De Ley, P., Garey, J.R., Liu, L.X., Scheldeman, P., Vierstraete, A., *et al.* (1998). A molecular evolutionary framework for the phylum Nematoda. *Nature*, 392, 71–75.
Boothby, T.C., Tenlen, J.R., Smith, F.W., Wang, J.R., Patanella, K.A., Nishimura, E.O., *et al.* (2015). Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proc. Natl. Acad. Sci. U. S. A.*, 112, 15976–15981.
Bourguinat, C., Ardelli, B.F., Pion, S.D.S., Kamgno, J., Gardon, J., Duke, B.O.L., *et al.* (2008). P-glycoprotein-like protein, a possible genetic marker for ivermectin resistance selection in Onchocerca volvulus. *Mol. Biochem. Parasitol.*, 158, 101–111.
Bradley, J.E., Nirmalan, N., Kläger, S.L., Faulkner, H. & Kennedy, M.W. (2001). River blindness: a role for parasite retinoid-binding proteins in the generation of pathology? *Trends Parasitol.*, 17, 471–475.
Braukmann, F., Jordan, D. & Miska, E.A. (2019). A genetic pathway encoding double-stranded RNA transporters and interactors regulates growth and plasticity in Caenorhabditis elegans. *bioRxiv*.
Brenner, S. (1974). The genetics of Caenorhabditis elegans. *Genetics*, 77, 71–94.
Brownlee, G.G., Sanger, F. & Barrell, B.G. (1967). Nucleotide sequence of 5S-ribosomal RNA from Escherichia coli. *Nature*, 215, 735–736.
Buchfink, B., Xie, C. & Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, 12, 59–60.
Bürglin, T.R., Lobos, E. & Blaxter, M.L. (1998). Caenorhabditis elegans as a model for parasitic nematodes. *Int. J. Parasitol.*, 28, 395–411.
Burke, M., Scholl, E.H., Bird, D.M., Schaff, J.E., Colman, S.D., Crowell, R., *et al.* (2015). The plant parasite Pratylenchus coffeae carries a minimal nematode genome. *Nematology*, 17, 621–637.
Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., *et al.* (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, 10, 421.
Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25, 1972–1973.
Cardona, J., González, M. & Álvarez, J. (2010). Otitis bovina por Rhabditis bovis en Córdoba, Colombia. Reporte de dos casos. *Revista MVZ Córdoba*, 15.
CAVALIER-SMITH & T. (1985). Cell volume and the evolution of eukaryote genome size. *The Evolution of Genome Size*, 105–184.
C. elegans Sequencing Consortium. (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science*, 282, 2012–2018.
Charlesworth, D. & Wright, S.I. (2001). Breeding systems and genome evolution. *Curr. Opin. Genet. Dev.*, 11, 685–690.
Chikhi, R. & Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, 30, 31–37.
Cho, S., Jin, S.-W., Cohen, A. & Ellis, R.E. (2004). A phylogeny of caenorhabditis reveals frequent loss of introns during nematode evolution. *Genome Res.*, 14, 1207–1220.
Coghlan, A. & Wolfe, K.H. (2004). Origins of recently gained introns in

Caenorhabditis. *Proc. Natl. Acad. Sci. U. S. A.*, 101, 11362–11367.
Compeau, P.E.C., Pevzner, P.A. & Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nat. Biotechnol.*, 29, 987–991.
Cook, D.E., Valle-Inclan, J.E., Pajoro, A., Rovenich, H., Thomma, B.P.H.J. & Faino, L. (2019a). Long-Read Annotation: Automated Eukaryotic Genome Annotation Based on Long-Read cDNA Sequencing. *Plant Physiol.*, 179, 38–54.
Cook, D.E., Zdraljevic, S., Roberts, J.P. & Andersen, E.C. (2017). CeNDR, the Caenorhabditis elegans natural diversity resource. *Nucleic Acids Res.*, 45, D650–D657.
Cook, S.J., Jarrell, T.A., Brittin, C.A., Wang, Y., Bloniarz, A.E., Yakovlev, M.A., *et al.* (2019b). Whole-animal connectomes of both Caenorhabditis elegans sexes. *Nature*, 571, 63–71.
Corsi, A.K., Wightman, B. & Chalfie, M. (2018). *A Transparent window into biology: A primer on Caenorhabditis elegans*. WormBook.
Cotton, J.A., Bennuru, S., Grote, A., Harsha, B., Tracey, A., Beech, R., *et al.* (2016). The genome of Onchocerca volvulus, agent of river blindness. *Nat Microbiol*, 2, 16216.
Coulombe-Huntington, J. & Majewski, J. (2007). Intron loss and gain in Drosophila. *Mol. Biol. Evol.*, 24, 2842–2850.
Cui, L., Wall, P.K., Leebens-Mack, J.H., Lindsay, B.G., Soltis, D.E., Doyle, J.J., *et al.* (2006). Widespread genome duplications throughout the history of flowering plants. *Genome Res.*, 16, 738–749.
Cutter, A.D., Baird, S.E. & Charlesworth, D. (2006). High nucleotide polymorphism and rapid decay of linkage disequilibrium in wild populations of Caenorhabditis remanei. *Genetics*, 174, 901–913.
Danecek, P., Schiffels, S. & Durbin, R. (2014). Multiallelic calling model in bcftools (-m).
Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M. & Blaxter, M.L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.*, 12, 499–510.
Dey, A., Chan, C.K.W., Thomas, C.G. & Cutter, A.D. (2013). Molecular hyperdiversity defines populations of the nematode Caenorhabditis brenneri. *Proc. Natl. Acad. Sci. U. S. A.*, 110, 11056–11060.
Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., *et al.* (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29, 15–21.
Drosophila 12 Genomes Consortium, Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., *et al.* (2007). Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, 450, 203–218.
Ellinghaus, D., Kurtz, S. & Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, 9, 18.
Emms, D.M. & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.*, 16, 157.
Fedorov, A., Roy, S., Fedorova, L. & Gilbert, W. (2003). Mystery of intron gain. *Genome Res.*, 13, 2236–2241.

Félix, M.-A. & Braendle, C. (2010). The natural history of Caenorhabditis elegans. *Curr. Biol.*, 20, R965–9.
Félix, M.-A., Braendle, C. & Cutter, A.D. (2014). A Streamlined System for Species Diagnosis in Caenorhabditis (Nematoda: Rhabditidae) with Name Designations for 15 Distinct Biological Species. *PLoS One*, 9, e94723.
Félix, M.-A. & Duveau, F. (2012). Population dynamics and habitat sharing of natural populations of Caenorhabditis elegans and C. briggsae. *BMC Biol.*, 10, 59.
Ferrari, C., Salle, R., Callemeyn-Torre, N., Jovelin, R., Cutter, A.D. & Braendle, C. (2017). Ephemeral-habitat colonization and neotropical species richness of Caenorhabditis nematodes. *BMC Ecol.*, 17, 43.
Ferraz, C.M., Sobral, S.A., Senna, C.C., Junior, O.F., Moreira, T.F., Tobias, F.L., *et al.* (2019). Combined use of ivermectin, dimethyl sulfoxide, mineral oil and nematophagous fungi to control Rhabditis spp. *Vet. Parasitol.*, 275, 108924.
Fierst, J.L., Willis, J.H., Thomas, C.G., Wang, W., Reynolds, R.M., Ahearne, T.E., *et al.* (2015). Reproductive Mode and the Evolution of Genome Size and Structure in Caenorhabditis Nematodes. *PLoS Genet.*, 11, e1005323.
Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E. & Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391, 806–811.
Fitch, D.H., Bugaj-Gaweda, B. & Emmons, S.W. (1995). 18S ribosomal RNA gene phylogeny for some Rhabditidae related to Caenorhabditis. *Mol. Biol. Evol.*, 12, 346–358.
Forst, S., Dowds, B., Boemare, N. & Stackebrandt, E. (1997). Xenorhabdus and Photorhabdus spp.: bugs that kill bugs. *Annu. Rev. Microbiol.*, 51, 47–72.
Fradin, H., Kiontke, K., Zegar, C., Gutwein, M., Lucas, J., Kovtun, M., *et al.* (2017). Genome Architecture and Evolution of a Unichromosomal Asexual Nematode. *Curr. Biol.*, 27, 2928–2939.e6.
Garofalo, A., Rowlinson, M.-C., Amambua, N.A., Hughes, J.M., Kelly, S.M., Price, N.C., *et al.* (2003). The FAR protein family of the nematode Caenorhabditis elegans. Differential lipid binding properties, structural characteristics, and developmental regulation. *J. Biol. Chem.*, 278, 8065–8074.
Garrison, E. & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv [q-bio.GN]*.
Gatesy, J. & Springer, M.S. (2014). Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Mol. Phylogenet. Evol.*, 80, 231–266.
Genomic classification of protein-coding gene families. (2019). . Available at: http://www.wormbook.org/chapters/www_genomclassprot/genomclassprot.html. Last accessed 8 December 2019.
Gilleard, J.S. (2004). The use of Caenorhabditis elegans in parasitic nematode research. *Parasitology*, 128 Suppl 1, S49–70.
Goodwin, S., McPherson, J.D. & McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, 17, 333–351.
Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29, 1072–1075.

Haas, B. (2007). TransposonPSI: an application of PSI-Blast to mine (retro-) transposon ORF homologies. *Broad Institute, Cambridge, MA, USA*.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., *et al.* (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.*, 8, 1494–1512.

Heled, J. & Drummond, A.J. (2010). Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.*, 27, 570–580.

Hillier, L.W., Miller, R.D., Baird, S.E., Chinwalla, A., Fulton, L.A., Koboldt, D.C., *et al.* (2007). Comparison of C. elegans and C. briggsae genome sequences reveals extensive conservation of chromosome organization and synteny. *PLoS Biol.*, 5, e167.

Hiraki, H., Kagoshima, H., Kraus, C., Schiffer, P.H., Ueta, Y., Kroiher, M., *et al.* (2017). Genome analysis of Diploscapter coronatus: insights into molecular peculiarities of a nematode with parthenogenetic reproduction. *BMC Genomics*, 18, 478.

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q. & Vinh, L.S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.*, 35, 518–522.

Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. (2016). BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, 32, 767–769.

Howe, K.L., Bolt, B.J., Shafie, M., Kersey, P. & Berriman, M. (2017). WormBase ParaSite - a comprehensive resource for helminth genomics. *Mol. Biochem. Parasitol.*, 215, 2–10.

Huerta-Cepas, J., Serra, F. & Bork, P. (2016a). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.*, 33, 1635–1638.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M.C., *et al.* (2016b). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.*, 44, D286–93.

Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., *et al.* (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.*, 36, 338–345.

Jain, M., Olsen, H.E., Paten, B. & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.*, 17, 239.

Janssen, I.J.I., Krücken, J., Demeler, J., Basiaga, M., Kornaś, S. & von Samson-Himmelstjerna, G. (2013a). Genetic variants and increased expression of Parascaris equorum P-glycoprotein-11 in populations with decreased ivermectin susceptibility. *PLoS One*, 8, e61635.

Janssen, I.J.I., Krücken, J., Demeler, J. & von Samson-Himmelstjerna, G. (2013b). Caenorhabditis elegans: modest increase of susceptibility to ivermectin in individual P-glycoprotein loss-of-function strains. *Exp. Parasitol.*, 134, 171–177.

Jeffroy, O., Brinkmann, H., Delsuc, F. & Philippe, H. (2006). Phylogenomics: the

beginning of incongruence? *Trends Genet.*, 22, 225–231.
Jiang, H., Lei, R., Ding, S.-W. & Zhu, S. (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, 15, 182.
Jones, D. & Candido, E.P. (1999). Feeding is inhibited by sublethal concentrations of toxicants and by heat stress in the nematode Caenorhabditis elegans: relationship to the cellular stress response. *J. Exp. Zool.*, 284, 147–157.
Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., *et al.* (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30, 1236–1240.
Junier, T. & Zdobnov, E.M. (2010). The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics*, 26, 1669–1670.
Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, 110, 462–467.
Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., *et al.* (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.*, 24, 1384–1395.
Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A. & Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods*, 14, 587–589.
Kanzaki, N., Tsai, I.J., Tanaka, R., Hunt, V.L., Liu, D., Tsuyama, K., *et al.* (2018). Biology and genome of a newly discovered sibling species of Caenorhabditis elegans. *Nat. Commun.*, 9, 3216.
Katoh, K. & Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, 30, 772–780.
Keller, M.W., Rambo-Martin, B.L., Wilson, M.M., Ridenour, C.A., Shepard, S.S., Stark, T.J., *et al.* (2018). Direct RNA Sequencing of the Coding Complete Influenza A Virus Genome. *Sci. Rep.*, 8, 14408.
Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E.S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423, 241–254.
Kidwell, M.G. (2002). Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, 115, 49–63.
Kim, J.-Y., Cho, M.K., Choi, S.H., Lee, K.H., Ahn, S.C., Kim, D.-H., *et al.* (2010). Inhibition of dextran sulfate sodium (DSS)-induced intestinal inflammation via enhanced IL-10 and TGF-β production by galectin-9 homologues isolated from intestinal parasites. *Mol. Biochem. Parasitol.*, 174, 53–61.
Kingan, S.B., Heaton, H., Cudini, J., Lambert, C.C., Baybayan, P., Galvin, B.D., *et al.* (2019). A High-Quality De novo Genome Assembly from a Single Mosquito Using PacBio Sequencing. *Genes* , 10.
Kiontke, K. (1997). Description of Rhabditis (Caenorhabditis) drosophilae n. sp. and R.(C.) sonorae n. sp.(Nematoda: Rhabditida) from saguaro cactus rot in Arizona. *Fundam. Appl. Nematol.*, 20, 305–315.
Kiontke, K.C., Félix, M.-A., Ailion, M., Rockman, M.V., Braendle, C., Pénigault, J.-B.,

et al. (2011). A phylogeny and molecular barcodes for Caenorhabditis, with numerous new species from rotting fruits. *BMC Evol. Biol.*, 11, 339.
Kiontke, K., Gavin, N.P., Raynes, Y., Roehrig, C., Piano, F. & Fitch, D.H.A. (2004). Caenorhabditis phylogeny predicts convergence of hermaphroditism and extensive intron loss. *Proc. Natl. Acad. Sci. U. S. A.*, 101, 9003–9008.
Kiontke, K. & Sudhaus, W. (2006). Ecology of Caenorhabditis species. *WormBook*, 1–14.
Kocot, K.M., Citarella, M.R., Moroz, L.L. & Halanych, K.M. (2013). PhyloTreePruner: A Phylogenetic Tree-Based Approach for Selection of Orthologous Sequences for Phylogenomics. *Evol. Bioinform. Online*, 9, 429–435.
Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P.A. (2018). Assembly of Long Error-Prone Reads Using Repeat Graphs. *bioRxiv*.
Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. & Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.*, 27, 722–736.
Koutsovoulos, G., Kumar, S., Laetsch, D.R., Stevens, L., Daub, J., Conlon, C., *et al.* (2016). No evidence for extensive horizontal gene transfer in the genome of the tardigrade Hypsibius dujardini. *Proc. Natl. Acad. Sci. U. S. A.*
Kozak, K.M., Owen McMillan, W., Joron, M. & Jiggins, C.D. (2018). Genome-wide admixture is common across the Heliconius radiation. *bioRxiv*.
Kreis, H.A. (1964). Ein neuer Nematode aus dem äusseren Gehörgang von Zeburindern in Ostafrika, Rhabditis bovis n. sp.(Rhabditidoidea; Rhabditidae).
Kreis, H.A. & Faust, E.C. (1933). Two New Species of Rhabditis (Rhabditis macrocerca and R. clavopapillata) Associated with Dogs and Monkeys in Experimental Strongyloides Studies. *Trans. Am. Microsc. Soc.*, 52, 162–172.
Kubatko, L.S. & Degnan, J.H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.*, 56, 17–24.
Laetsch, D.R. & Blaxter, M.L. (2017a). BlobTools: Interrogation of genome assemblies. *F1000Res.*, 6.
Laetsch, D.R. & Blaxter, M.L. (2017b). KinFin: Software for Taxon-Aware Analysis of Clustered Protein Sequences. *G3* , 7, 3349–3357.
Lamelza, P. & Ailion, M. (2016). Cytoplasmic-nuclear incompatibility between wild-isolates of Caenorhabditis nouraguensis. *bioRxiv*.
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, 409, 860–921.
Lartillot, N. & Philippe, H. (2004). A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. *Mol. Biol. Evol.*, 21, 1095–1109.
Lartillot, N., Rodrigue, N., Stubbs, D. & Richer, J. (2013). PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.*, 62, 611–615.
Lasken, R.S. & Stockwell, T.B. (2007). Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol.*, 7, 19.
Lee, R.Y.N., Howe, K.L., Harris, T.W., Arnaboldi, V., Cain, S., Chan, J., *et al.* (2018). WormBase 2017: molting into a new stage. *Nucleic Acids Res.*, 46, D869–D874.

Letunic, I. & Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.*, 44, W242–5.
Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]*.
Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34, 3094–3100.
Llorens, C., Futami, R., Covelli, L., Domínguez-Escribá, L., Viu, J.M., Tamarit, D., *et al.* (2011). The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Res.*, 39, D70–4.
Loman, N.J., Quick, J. & Simpson, J.T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods*, 12, 733–735.
Lynch, M. (2002). Intron evolution as a population-genetic process. *Proc. Natl. Acad. Sci. U. S. A.*, 99, 6118–6123.
Lynch, M. & Conery, J.S. (2003). The origins of genome complexity. *Science*, 302, 1401–1404.
Marçais, G. & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27, 764–770.
Martin, S.H., Dasmahapatra, K.K., Nadeau, N.J., Salazar, C., Walters, J.R., Simpson, F., *et al.* (2013). Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Res.*, 23, 1817–1828.
Maupas, E. (1900). *Modes et formes de reproduction des nematodes*.
Miller, D.E., Staber, C., Zeitlinger, J. & Hawley, R.S. (2018). Highly Contiguous Genome Assemblies of 15 Drosophila Species Generated Using Nanopore Sequencing. *G3* , 8, 3131–3141.
Milstone, A.M., Harrison, L.M., Bungiro, R.D., Kuzmic, P. & Cappello, M. (2000). A broad spectrum Kunitz type serine protease inhibitor secreted by the hookworm Ancylostoma ceylanicum. *J. Biol. Chem.*, 275, 29391–29399.
Minh, B.Q., Hahn, M.W. & Lanfear, R. (2018). New methods to calculate concordance factors for phylogenomic datasets. *bioRxiv*.
Mohamed, S. & Syed, B.A. (2013). Commercial prospects for genomic sequencing technologies. *Nat. Rev. Drug Discov.*, 12, 341–342.
Mortazavi, A., Schwarz, E.M., Williams, B., Schaeffer, L., Antoshechkin, I., Wold, B.J., *et al.* (2010). Scaffolding a Caenorhabditis nematode genome with RNA-seq. *Genome Res.*, 20, 1740–1747.
Msolla, P., Falmer-Hansen, J., Musemakweli, J. & Monrad, J. (1985). Treatment of bovine parasitic otitis using ivermectin. *Trop. Anim. Health Prod.*, 17, 166–168.
Msolla, P., Kimera, I.S., Kassuku, A.A. & Semuguruka, W.D. (1989). The role of flies, manure and soil in the epidemiology of bovine parasitic otitis. *Proc. 7th Tanzania Vet*.
Msolla, P., Semuguruka, W.D., Kasuku, A.A. & Shoo, M.K. (1993). Clinical observations on bovine parasitic otitis in Tanzania. *Trop. Anim. Health Prod.*, 25, 15–18.
Naville, M., Henriet, S., Warren, I., Sumic, S., Reeve, M., Volff, J.-N., *et al.* (2019). Massive Changes of Genome Size Driven by Expansions of Non-autonomous

Transposable Elements. *Curr. Biol.*, 29, 1161–1168.e6.
Nguyen, L.-T., Schmidt, H.A., von Haeseler, A. & Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, 32, 268–274.
Nowell, R.W., Almeida, P., Wilson, C.G., Smith, T.P., Fontaneto, D., Crisp, A., *et al.* (2018). Comparative genomics of bdelloid rotifers: Insights from desiccating and nondesiccating species. *PLoS Biol.*, 16, e2004830.
Nowoshilow, S., Schloissnig, S., Fei, J.-F., Dahl, A., Pang, A.W.C., Pippel, M., *et al.* (2018). The axolotl genome and the evolution of key tissue formation regulators. *Nature*, 554, 50–55.
Orme, D., Freckleton, R., Thomas, G. & Petzoldt, T. (2013). The caper package: comparative analysis of phylogenetics and evolution in R. *R package version*, 5, 1–36.
Paradis, E. & Schliep, K. (2018). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*.
Pennell, M.W., Eastman, J.M., Slater, G.J., Brown, J.W., Uyeda, J.C., FitzJohn, R.G., *et al.* (2014). geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics*, 30, 2216–2218.
Petrov, D.A. (2001). Evolution of genome size: new approaches to an old problem. *Trends Genet.*, 17, 23–28.
Picard Tools - By Broad Institute. (2019). . Available at: https://broadinstitute.github.io/picard/. Last accessed 8 December 2019.
Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., Team, R.C. & Others. (2012). nlme: Linear and nonlinear mixed effects models. *R package version*, 3.
Platt, R.N., 2nd, Vandewege, M.W. & Ray, D.A. (2018). Mammalian transposable elements and their impacts on genome evolution. *Chromosome Res.*, 26, 25–43.
Pundir, S., Martin, M.J. & O'Donovan, C. (2017). UniProt Protein Knowledgebase. *Methods Mol. Biol.*, 1558, 41–55.
Rambaut, A., Drummond, A.J. & Suchard, M. (2007). Tracer v1. 6.
Revell, L.J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.*
Richaud, A., Zhang, G., Lee, D., Lee, J. & Félix, M.-A. (2018). The Local Coexistence Pattern of Selfing Genotypes in Caenorhabditis elegans Natural Metapopulations. *Genetics*, 208, 807–821.
Riddle, D.L., Blumenthal, T., Meyer, B.J. & Priess, J.R. (Eds.). (2011). *C. elegans II*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY).
Robertson, H.M. (1998). Two large families of chemoreceptor genes in the nematodes Caenorhabditis elegans and Caenorhabditis briggsae reveal extensive gene duplication, diversification, movement, and intron loss. *Genome Res.*, 8, 449–463.
Roberts, R.J., Carneiro, M.O. & Schatz, M.C. (2013). The advantages of SMRT sequencing. *Genome Biol.*, 14, 405.
Roch, S. & Warnow, T. (2015). On the Robustness to Gene Tree Estimation Error (or lack thereof) of Coalescent-Based Species Tree Methods. *Syst. Biol.*, 64, 663–676.

Rockman, M.V. & Kruglyak, L. (2009). Recombinational landscape and population genomics of Caenorhabditis elegans. *PLoS Genet.*, 5, e1000419.
Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, 4, e2584.
Rokas, A. & Holland, P.W. (2000). Rare genomic changes as a tool for phylogenetics. *Trends Ecol. Evol.*, 15, 454–459.
Rokas, A., Williams, B.L., King, N. & Carroll, S.B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425, 798–804.
Roy, S.W., Fedorov, A. & Gilbert, W. (2003). Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc. Natl. Acad. Sci. U. S. A.*, 100, 7158–7162.
Roy, S.W. & Penny, D. (2006). Smoke without fire: most reported cases of intron gain in nematodes instead reflect intron losses. *Mol. Biol. Evol.*, 23, 2259–2262.
Roy, S.W. & Penny, D. (2007). Patterns of intron loss and gain in plants: intron loss-dominated evolution and genome-wide comparison of O. sativa and A. thaliana. *Mol. Biol. Evol.*, 24, 171–181.
Ruan, J. & Li, H. (2019). Fast and accurate long-read assembly with wtdbg2. *bioRxiv*.
Sabina, J. & Leamon, J.H. (2015). Bias in Whole Genome Amplification: Causes and Considerations. *Methods Mol. Biol.*, 1347, 15–41.
Sanderson, M.J., Purvis, A. & Henze, C. (1998). Phylogenetic supertrees: Assembling the trees of life. *Trends Ecol. Evol.*, 13, 105–109.
Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., *et al.* (1977a). Nucleotide sequence of bacteriophage ϕX174 DNA. *Nature*, 265, 687–695.
Sanger, F., Nicklen, S. & Coulson, A.R. (1977b). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.*, 74, 5463–5467.
Sarkies, P., Selkirk, M.E., Jones, J.T., Blok, V., Boothby, T., Goldstein, B., *et al.* (2015). Ancient and novel small RNA pathways compensate for the loss of piRNAs in multiple independent nematode lineages. *PLoS Biol.*, 13, e1002061.
Scheiber, S.H. (1880). Ein Fall von mikroskopisch kleinen Rundwürmern—Rhabditis genitalis—im Urin einer Kranken. *Virchows Arch.*, 82, 161–175.
Schirmer, M., Ijaz, U.Z., D'Amore, R., Hall, N., Sloan, W.T. & Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.*, 43, e37.
Schmidt, G. & Kuntz, R.E. (1972). Caenorhabditis avicola sp. n.(Rhabditidae) found in a bird from Taiwan. *Proc. Helminthol. Soc. Wash.*, 39, 189–191.
Sedlazeck, F.J., Lee, H., Darby, C.A. & Schatz, M.C. (2018). Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.*, 19, 329–346.
Sheps, J.A., Ralph, S., Zhao, Z., Baillie, D.L. & Ling, V. (2004). The ABC transporter gene family of Caenorhabditis elegans has implications for the evolutionary dynamics of multidrug resistance in eukaryotes. *Genome Biol.*, 5, R15.
Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.*, 51, 492–508.
Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with

single-copy orthologs. *Bioinformatics*, 31, 3210–3212.
Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M. & Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Res.*, 19, 1117–1123.
Slos, D., Sudhaus, W., Stevens, L., Bert, W. & Blaxter, M. (2017). Caenorhabditis monodelphis sp. n.: defining the stem morphology and genomics of the genus Caenorhabditis. *BMC Zoology*, 2, 4.
Smit, A.F.A., Hubley, R. & Green, P. (1996). *RepeatMasker*. Available at: http://www.repeatmasker.org/. Last accessed .
Smit, A. & Hubley, R. (2010). *RepeatModeler Open-1.0*. Available at: http://www.repeatmasker.org/RepeatModeler/. Last accessed .
Snutch, T.P. & Baillie, D.L. (1983). Alterations in the pattern of gene expression following heat shock in the nematode Caenorhabditis elegans. *Can. J. Biochem. Cell Biol.*, 61, 480–487.
Spieth, J., Brooke, G., Kuersten, S., Lea, K. & Blumenthal, T. (1993). Operons in C. elegans: polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. *Cell*, 73, 521–532.
Stark, A., Kheradpour, P., Parts, L., Brennecke, J., Hodges, E., Hannon, G.J., *et al.* (2007). Systematic discovery and characterization of fly microRNAs using 12 Drosophila genomes. *Genome Res.*, 17, 1865–1879.
Steinbiss, S., Willhoeft, U., Gremme, G. & Kurtz, S. (2009). Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Res.*, 37, 7002–7013.
Stein, L.D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M.R., Chen, N., *et al.* (2003). The Genome Sequence of Caenorhabditis briggsae: A Platform for Comparative Genomics. *PLoS Biol.*, 1, e45.
Stevens, L., Félix, M.-A., Beltran, T., Braendle, C., Caurcel, C., Fausett, S., *et al.* (2019). Comparative genomics of 10 new Caenorhabditis species. *Evolution Letters*, 3, 217–236.
Stricklin, S.L., Griffiths-Jones, S. & Eddy, S.R. (2005). C. elegans noncoding RNA genes. *WormBook*, 1–7.
Sudhaus, W. (1974). *Zur Systematik, Verbreitung, Ökologie und Biologie neuer und wenig bekannter Rhabditiden (Nematoda)*.
Sudhaus, W. & Kiontke, K. (1996). Phylogeny of Rhabditis subgenus Caenorhabditis (Rhabditidae, Nematoda)*. *J. Zoolog. Syst. Evol. Res.*, 34, 217–233.
Sulston, J.E. & Brenner, S. (1974). The DNA of Caenorhabditis elegans. *Genetics*, 77, 95–104.
Sulston, J.E., Schierenberg, E., White, J.G. & Thomson, J.N. (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev. Biol.*, 100, 64–119.
Suyama, M., Torrents, D. & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.*, 34, W609–12.
Sverdlov, A.V., Babenko, V.N., Rogozin, I.B. & Koonin, E.V. (2004). Preferential loss and gain of introns in 3' portions of genes suggests a reverse-transcription mechanism of intron insertion. *Gene*, 338, 85–91.
Takeuchi, T., Kawashima, T., Koyanagi, R., Gyoja, F., Tanaka, M., Ikuta, T., *et al.*

(2012). Draft genome of the pearl oyster Pinctada fucata: a platform for understanding bivalve biology. *DNA Res.*, 19, 117–130.
Tandonnet, S., Koutsovoulos, G.D., Adams, S., Cloarec, D., Parihar, M., Blaxter, M.L., *et al.* (2019). Chromosome-Wide Evolution and Sex Determination in the Three-Sexed Nematode Auanema rhodensis. *G3* , 9, 1211–1230.
Tan, L. & Grewal, P.S. (2002). Endotoxin activity of Moraxella osloensis against the grey garden slug, Deroceras reticulatum. *Appl. Environ. Microbiol.*, 68, 3943–3947.
Tarailo-Graovac, M. & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics*, 25, 4–10.
Taubert, S., Ward, J.D. & Yamamoto, K.R. (2011). Nuclear hormone receptors in nematodes: evolution and function. *Mol. Cell. Endocrinol.*, 334, 49–55.
Team, R.C. (2018). R: A language and environment for statistical computing; 2015.
Tenaillon, M.I., Hufford, M.B., Gaut, B.S. & Ross-Ibarra, J. (2011). Genome size and transposable element content as determined by high-throughput sequencing in maize and Zea luxurians. *Genome Biol. Evol.*, 3, 219–229.
Thomas, C.G., Li, R., Smith, H.E., Woodruff, G.C., Oliver, B. & Haag, E.S. (2012). Simplification and desexualization of gene expression in self-fertile nematodes. *Curr. Biol.*, 22, 2167–2172.
Thomma, B.P.H.J., Seidl, M.F., Shi-Kunne, X., Cook, D.E., Bolton, M.D., van Kan, J.A.L., *et al.* (2016). Mind the gap; seven reasons to close fragmented genome assemblies. *Fungal Genet. Biol.*, 90, 24–30.
Turner, D.G., Wildblood, L.A., Inglis, N.F. & Jones, D.G. (2008). Characterization of a galectin-like activity from the parasitic nematode, Haemonchus contortus, which modulates ovine eosinophil migration in vitro. *Vet. Immunol. Immunopathol.*, 122, 138–145.
Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.*, 27, 737–746.
Verocai, G.G., Fernandes, J.I., Correia, T.R., Melo, R.M., Alves, P.A.M., Scott, F.B., *et al.* (2009). Inefficacy of albendazole sulphoxide and ivermectin for the treatment of bovine parasitic otitis caused by rhabditiform nematodes. *Pesqui. Vet. Bras.*, 29, 910–912.
Vinson, J.P., Jaffe, D.B., O'Neill, K., Karlsson, E.K., Stange-Thomann, N., Anderson, S., *et al.* (2005). Assembly of polymorphic genomes: algorithms and application to Ciona savignyi. *Genome Res.*, 15, 1127–1135.
Volk, J. & Others. (1950). Nematodes associated with earthworms and carrion beetles. *Zoologische Jahrbucher. Abteilung fur Systematik, okologie und Geographie der Tiere.*, 79, 1–70.
Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J., *et al.* (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*, 33, 2202–2204.
Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., *et al.* (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, 9, e112963.
Walton, A.C. (1959). Some parasites and their chromosomes. *J. Parasitol.*, 45, 1–20.
Wang, B., Tseng, E., Regulski, M., Clark, T.A., Hon, T., Jiao, Y., *et al.* (2016).

Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.*, 7, 11708.
Wang, W., Wang, S., Zhang, H., Yuan, C., Yan, R., Song, X., *et al.* (2014). Galectin Hco-gal-m from Haemonchus contortus modulates goat monocytes and T cell function in different patterns. *Parasit. Vectors*, 7, 342.
Watson, M. & Warr, A. (2019). Errors in long-read assemblies can critically affect protein prediction. *Nat. Biotechnol.*, 37, 124–126.
White, J.G., Southgate, E., Thomson, J.N. & Brenner, S. (1986). The structure of the nervous system of the nematode Caenorhabditis elegans. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 314, 1–340.
Wickham, H. (2009). *Ggplot2: Elegant Graphics for Data Analysis*. 2nd edn. Springer Publishing Company, Incorporated.
Wick, R.R., Judd, L.M. & Holt, K.E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. *bioRxiv*.
Xu, M., Molento, M., Blackhall, W., Ribeiro, P., Beech, R. & Prichard, R. (1998). Ivermectin resistance in nematodes may be caused by alteration of P-glycoprotein homolog. *Mol. Biochem. Parasitol.*, 91, 327–335.
Yang, H.-P. & Barbash, D.A. (2008). Abundant and species-specific DINE-1 transposable elements in 12 Drosophila genomes. *Genome Biol.*, 9, R39.
Yenerall, P. & Zhou, L. (2012). Identifying the mechanisms of intron gain: progress and trends. *Biol. Direct*, 7, 29.
Yin, D., Schwarz, E.M., Thomas, C.G., Felde, R.L., Korf, I.F., Cutter, A.D., *et al.* (2018). Rapid genome shrinkage in a self-fertile nematode reveals sperm competition proteins. *Science*, 359, 55–61.
Yoshida, Y., Koutsovoulos, G., Laetsch, D.R., Stevens, L., Kumar, S., Horikawa, D.D., *et al.* (2017). Comparative genomics of the tardigrades Hypsibius dujardini and Ramazzottius varieornatus. *PLoS Biol.*, 15, e2002266.
You, M., Yue, Z., He, W., Yang, X., Yang, G., Xie, M., *et al.* (2013). A heterozygous moth genome provides insights into herbivory and detoxification. *Nat. Genet.*, 45, 220–225.
Zerbino, D.R. & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, 18, 821–829.
Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, 19, 153.
Zimmermann, T., Mirarab, S. & Warnow, T. (2014). BBCA: Improving the scalability of* BEAST using random binning. *BMC Genomics*, 15, S11.
Zulkapli, M.M., Rosli, M.A.F., Salleh, F.I.M., Noor, N.M., Aizat, W.M. & Goh, H.-H. (2017). Iso-Seq analysis of Nepenthes ampullaria, Nepenthes rafflesiana and Nepenthes× hookeriana for hybridisation study in pitcher plants. *Genomics data*, 12, 130.

Appendix A: Supplementary materials for chapter 2

Table S1: Strains and author contributions

Species	*Strain*	*Inbred (Y/N)*	*DNA extraction*	*Sequencing**	*Assembly*	*Gene Prediction*
C. afra	JU1286	Y	AR		MN	LS
C. astrocarya	NIC1040	Y	AR		LS	LS
C. bovis	-	N	LS	LS	LS	LS
C. castelli	JU1956	Y	AR		LS	LS
C. dolens	NIC394	Y	AR		LS	LS
C. doughertyi	JU1771	Y	AR		SC	LS
C. drosophilae	DF5077	Y	AR		LS	LS
C. guadoupensis	NIC113	N	LS, AR		LS	LS
C. imperialis	EG5942	Y	LS		LS	LS
C. macrosperma	JU2083	Y	AR		LS	LS
C. monodelphis	JU1667	Y	LS, AR	LS	LS	LS
C. nouraguensis	JU2079	Y	AR		JY	LS
C. oiwi	ECA1100	Y	TC		LS	LS
C. parvicauda	NIC534	Y	AR		LS	LS
C. plicata	SB355	Y	AR		LS	LS
C. portoensis	EG5942	Y	AR		LS	LS
C. quiockensis	JU2809	Y	AR		CC	LS
***C.* sp. 2**	DF5070	Y	KK		LS	LS
***C.* sp. 24**	DF5173	Y	AR		LS	LS
***C.* sp. 25**	QG555	Y	AR		LS	LS
***C.* sp. 27**	ZF1457	Y	TC		LS	LS
***C.* sp. 30**	ECA211	Y	KK		LS	LS
***C.* sp. 46**	NIC1120	N	LS		LS	LS
***C.* sp. 48**	BRC20454	Y	LS		LS	LS
***C.* sp. 49**	BRC20456	N	LS		LS	LS
***C.* sp. 51**	QG2939	Y	LS		LS	LS
***C.* sp. 54**	BRC20483	Y	LS		LS	LS
***C.* sp. 55**	JU2215	Y	AR		LS	LS
***C.* sp. 56**	JU2215	Y	AR		LS	LS
***C.* sp. 8**	JU2788	Y	AR		LS	LS
C. sulstoni	JU2818	Y	AR		TB	LS
C. tribulationis	JU2585	Y	AR		LS	LS
C. uteleia	JU1968	Y	AR		LS	LS
C. virilis	NIC1070	N	AR		GK	LS
C. vivipara	NIC564	N	LS, CB	LS	LS	LS
C. waitukubuli	JU1898	Y	CB		LS	LS
C. wallacei	JU2190	Y	AR		LS	LS
C. zanzibari	NIC1040	Y	AR		LS	LS

Information regarding strain isolation and inbreeding can be found at rhabditina.org. Abbreviations: **AR:** Aurelien Richaud, **CB:** Christian Braendle, **CC:** Carlos Caurcel, **EG:** Edinburgh Genomics, **EM:** Eunice Machuka, **GK:** Georgious Koutsvolous, **JY:** Janet Young, , **KK:** Karin Kiontke, **LS:** Lewis Stevens, **MN:** Matthew Newton, **SC:** Sinduja Chandrasekar, **TB:** Toni Beltran, **TC:** Timothy Crombie. * All Illumina sequencing and PromethION sequencing was performed by Edinburgh genomics. **LS** performed minION sequencing.

Table S2: Assembly software used for each species

Species	Assembler	Version
C. afra	Velvet	1.2.10
C. astrocarya	SPAdes	3.1.1
C. bovis	wtdbg2	2.3
C. castelli	SPAdes	3.1.1
C. dolens	SPAdes	3.1.1
C. doughertyi	ABySS	1.9.0
C. drosophilae	SPAdes	3.1.1
C. guadeloupensis	wtdbg2	2.3
C. imperialis	SPAdes	3.1.1
C. macrosperma	SPAdes	3.1.1
C. monodelphis	flye	2.4.2
C. nouraguensis	Platanus	1.2.4
C. oiwi	Platanus	1.2.4
C. parvicauda	Platanus	1.2.4
C. plicata	SPAdes	3.1.1
C. portoensis	flye	2.4.2
C. quiockensis	Velvet	1.2.10
***C.* sp. 2**	SPAdes	3.1.1
***C.* sp. 8**	SPAdes	3.1.1
***C.* sp. 24**	SPAdes	3.1.1
***C.* sp. 25**	SPAdes	3.1.1
***C.* sp. 27**	SPAdes	3.1.1
***C.* sp. 30**	SPAdes	3.1.1
***C.* sp. 46**	Platanus	1.2.4
***C.* sp. 48**	Platanus	1.2.4
***C.* sp. 49**	Platanus	1.2.4
***C.* sp. 51**	Platanus	1.2.4
***C.* sp. 54**	Platanus	1.2.4
***C.* sp. 55**	SPAdes	3.1.1
***C.* sp. 56**	Platanus	1.2.4
C. sulstoni	Velvet	1.2.10
C. tribulationis	Velvet	1.2.10
C. uteleia	Velvet	1.2.10
C. virilis	SPAdes	3.1.1
C. vivipara	wtdbg2	2.3
C. waitukubuli	Platanus	1.2.4
C. wallacei	SPAdes	3.1.1
C. zanzibari	Platanus	1.2.4

Table S3: Software versions and relevant parameters

Software	*Version*	*Relevant parameters*
FastQC	v0.11.7	
Skewer	0.2.2	-q 30 -l 76 -m any
Blobtools	v1.0	
SPAdes	3.1.1	--only-assembler
NCBI-BLAST+	2.7.1+	-max_target_seqs 1 -max_hsps 1 -evalue 1e-25
Diamond	v0.9.17	--max-target-seqs 1 --evalue 1e-25
KmerGenie	1.7048	
JellyFish	2.2.7	-C -m 21
BUSCO	3.0.2	
SCUBAT2	-	
Guppy	v3.0.3	
Albacore	2.3.4	
Nanopolish	0.10.2	
Medaka	0.7.0	
Pilon	1.23	--fix bases
BWA	0.7.17-r1188	
Racon	v1.2.1	-m 8 -x -6 -g -8 -w 500
minimap2	2.13-r850	-ax map-ont
RepeatModeler	open-1.0.11	-engine ncbi
RepeatMasker	open-4.0.9	
BRAKER	2	--epmod
Trinity	2.6.5	

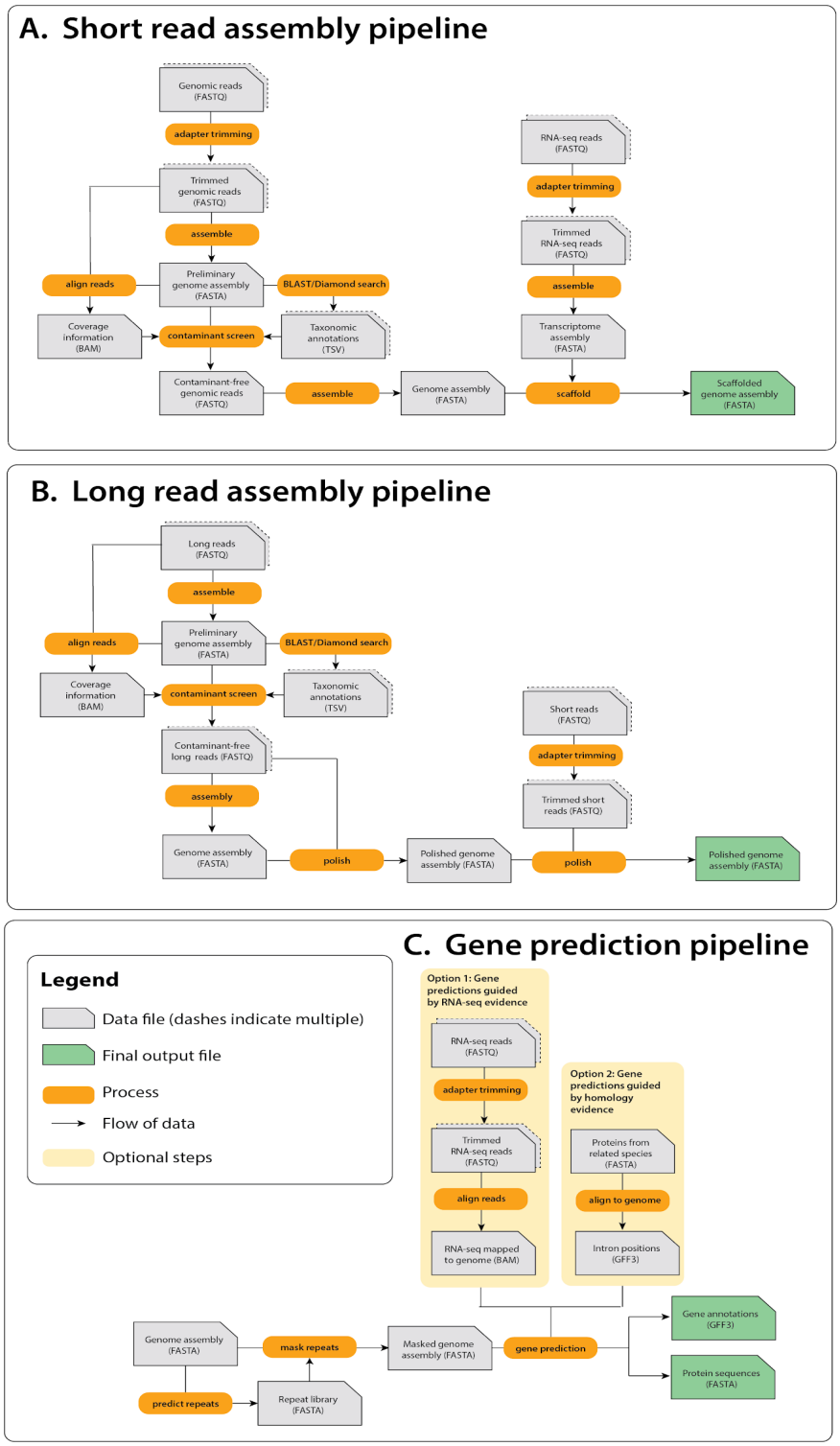

Figure S1: Genome assembly and gene prediction pipelines.

A: Short-read assembly pipeline. **B:** Long-read assembly pipeline. **C:** Protein-coding gene prediction pipeline.

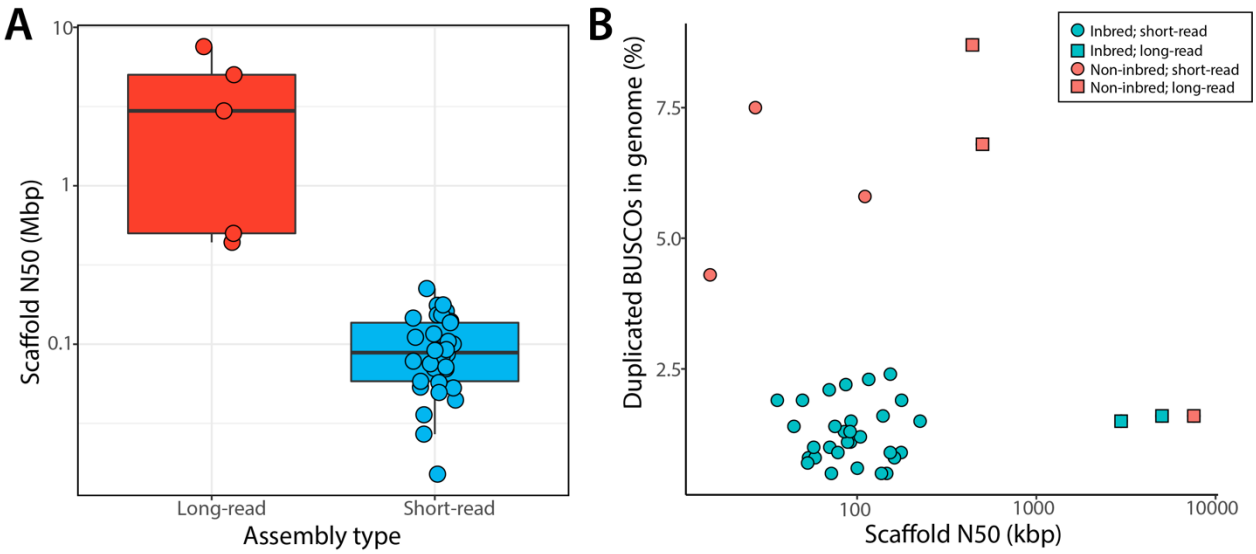

Figure S2: Effect of read length and heterozygosity on assembly contiguity. A: Scaffold N50s of long-read and short-read assemblies. **B:** Scaffold N50s and proportion of duplicated BUSCOs in each species.

Appendix B: Supplementary materials for chapter 3

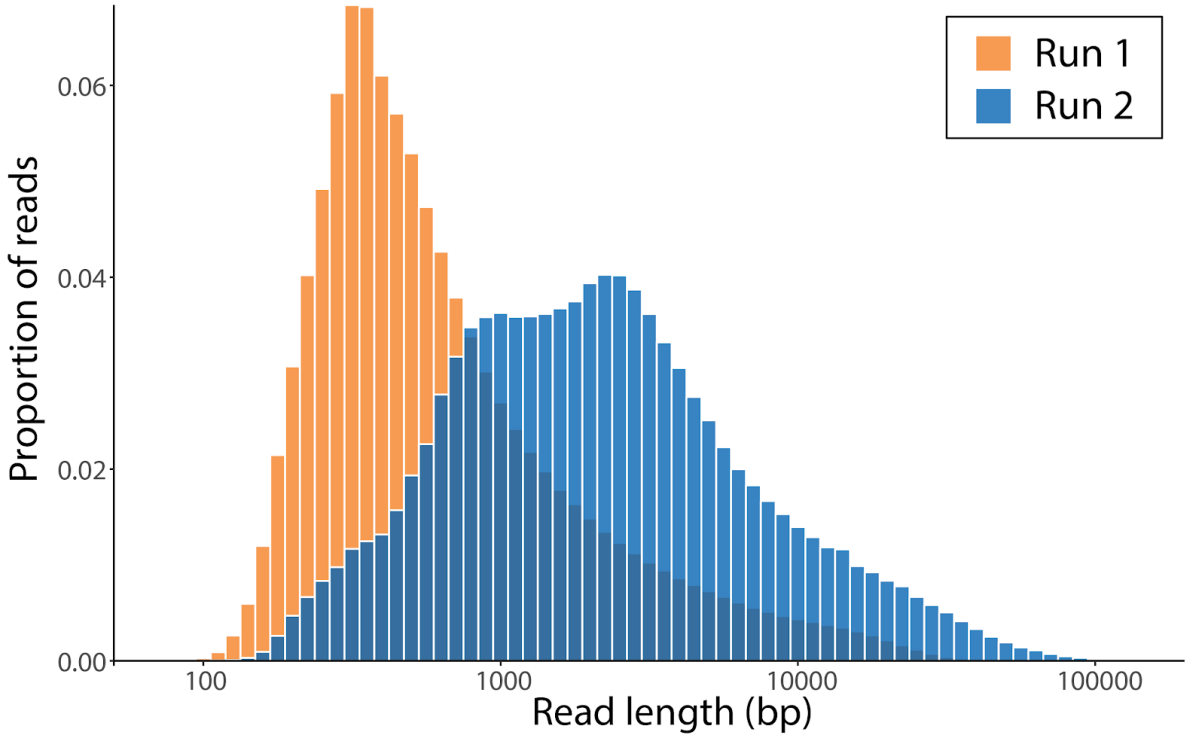

Figure S1: MinION read length histogram. Related to Table 1.
For Run 2, we prepared the library using the SQK-LS109 short fragment buffer (Oxford Nanopore) as the DNA appeared to be highly fragmented.

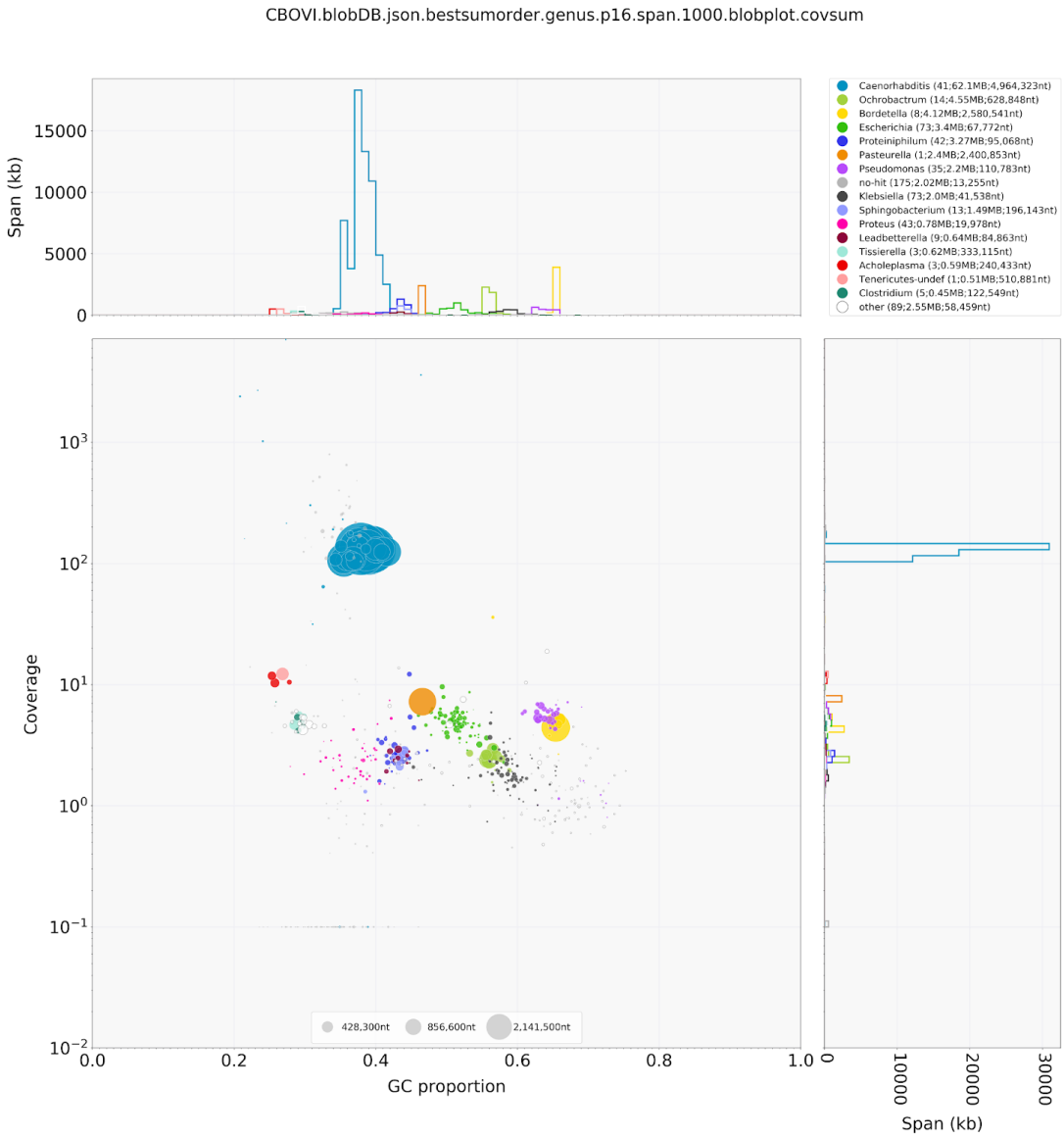

Figure S2: Taxon-annotation GC-coverage plot of a preliminary assembly of the *C. bovis* genome. Related to Table 1.
Results of NCBI-BLAST+ and Diamond searches of NCBI nucleotide 'nt' or UniProt Reference Proteomes databases were provided to blobtools which assigned taxonomy (using the 'bestsumorder' taxonomy rule). Coverage of each contig in the MinION read set is shown. Several contigs have top hits to bacterial species which are known mammalian pathogens, including *Pasteurella multocida* (cause of haemorrhagic septicaemia in cattle [96]), *Ochrobactrum anthropi,* and *Bordetella petrii* (both of which have been associated with opportunistic infections in humans [97,98]).

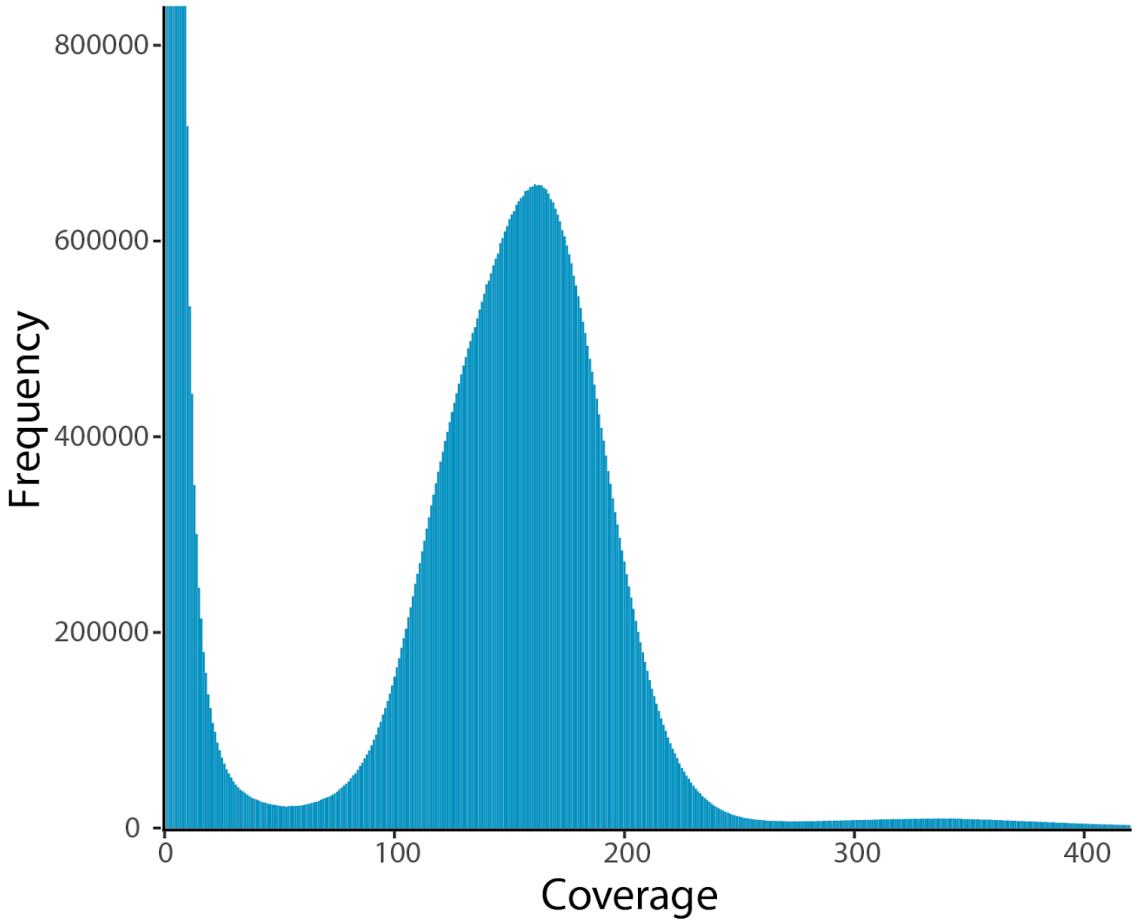

Figure S3: Kmer spectra of Illumina short-reads. Related to Table 1.
Kmers of length 19 were counted in Illumina MiSeq reads using Jellyfish. GenomeScope estimated 0.126% heterozygosity and a genome size of 61.3 Mb.

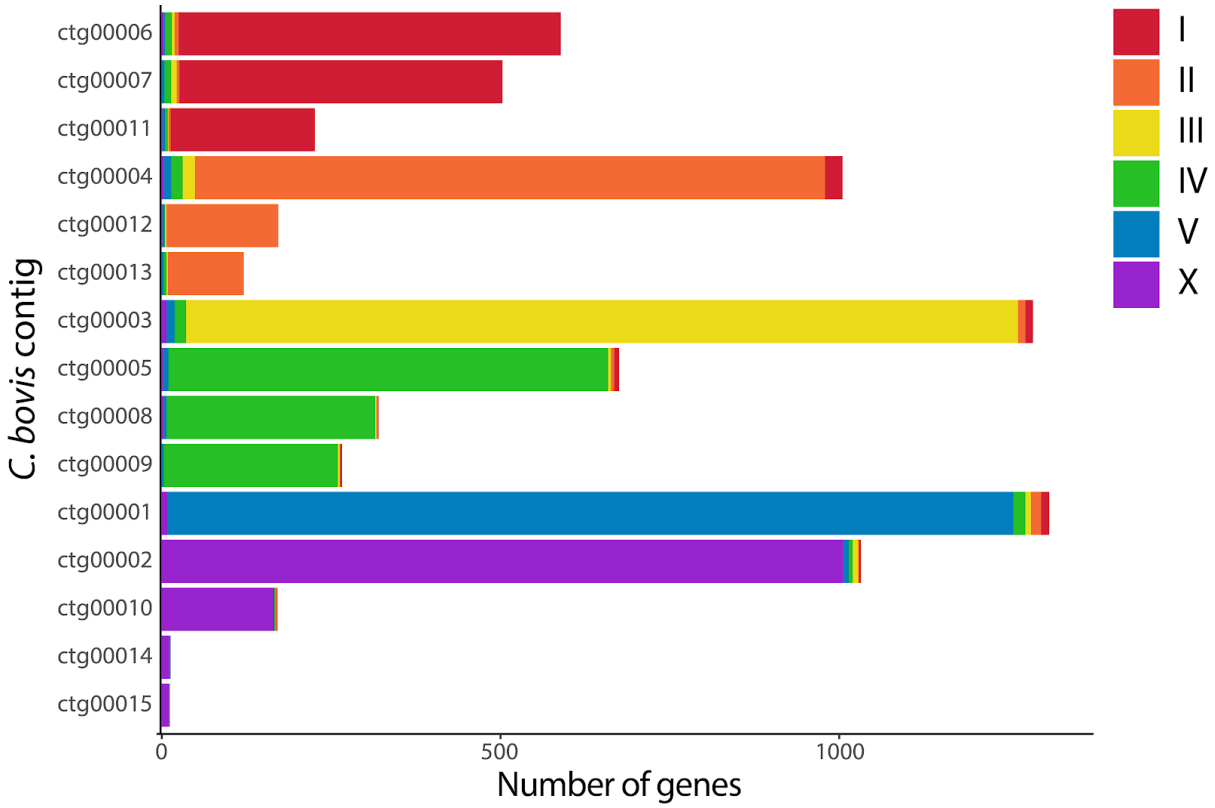

Figure S4: Composition of 15 *C. bovis* contigs. Related to Figure 2.
15 contigs (representing 99.4% of the *C. bovis* assembly) are shown. Bars represent the number of *C. bovis* genes with a *C. elegans* orthologue on each contig; bars are coloured by the chromosome location of the *C. elegans* orthologue. 20 contigs, each containing fewer than 10 *C. elegans* orthologues and cumulatively spanning 0.39 Mb, are not shown.

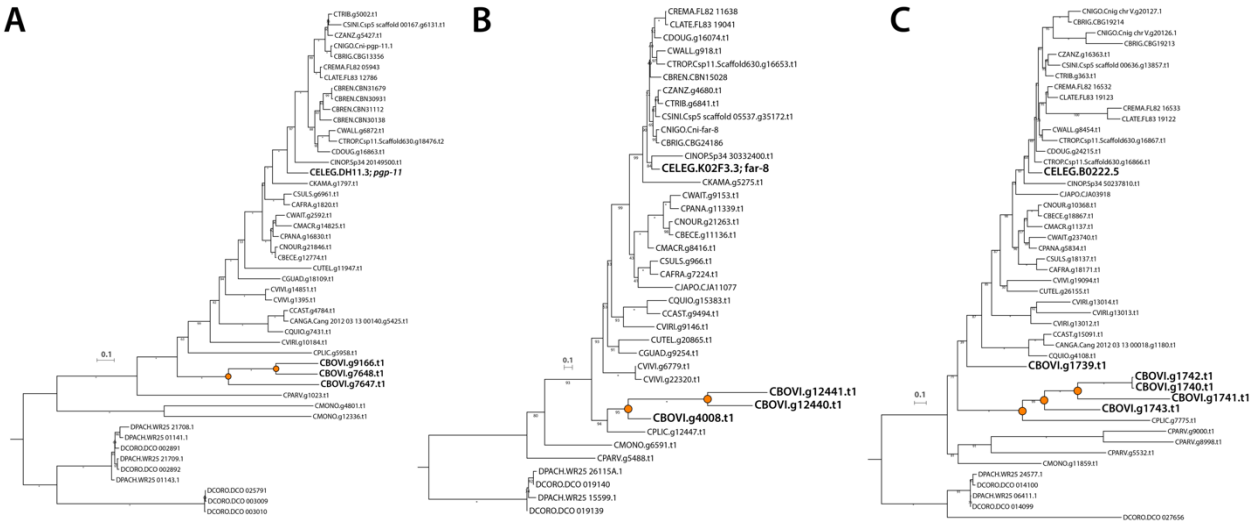

Figure S5: Expanded gene families in *C. bovis*. Related to Figure 4.
Maximum likelihood gene tree of the orthogroup containing (**A**) *pgp-11* (OG0002589) inferred using the LG+I+G4 substitution model, (**B**) *far-8* (OG0006033) inferred using the LG+G4 substitution model, and (**C**) containing a family of serine protease inhibitors (OG0002693) inferred using the WAG+I+G4 substitution model. *C. bovis* duplications are indicated by orange circles on internal nodes. Branch lengths represent the number of substitutions per site; scale is shown for each tree. Bootstrap values are displayed as branch annotations; "*" = 100.

Site	*Name*	*GPS coordinates*	*Type*	*Date*	*Number of animals sampled*
1	Amukura	0.55662731, 34.26680546	SH	26/03/2019	3
2	Kimilli	0.78995062, 34.72008678	SH	28/03/2019	7
3	Ikolomani	0.203020438, 34.66886937	LM	02/04/2019	7
4	Myanga	0.561452038, 34.38950072	LM	03/04/2019	7
5	Angurai	0.637058568, 34.2727974	LM	03/04/2019	7
6	Malaba	0.63711942, 34.27281931	SH	03/04/2019	3
7	Chwele	0.739193074, 34.58697939	LM	08/04/2019	7
8	Chwele	0.73913342, 34.58692581	SH	08/04/2019	2

Table S1: Sampling locations and number of animals sampled. Related to Figure 1.
LM: Livestock market, **SH**: slaughterhouse. *C. bovis* was isolated from an individual sampled at a livestock market in Chwele (site 7).

	Run 1	*Run 2*
Total data (Gbp)	4.42	6.83
Read count	955,868	4,857,912
Read length N50 (kbp)	11.41	4.32
Mean read length (kbp)	4.62	1.41
Longest read (kbp)	242.02	172.39

Table S2: MinION sequencing statistics. Related to Figure 2.
The library for 'Run 2' was prepared using the SQK-LS109 short fragment buffer (Oxford Nanopore) as the DNA appeared to be highly fragmented.

Gene family	Count in *C. elegans*	Count in *C. bovis*	Difference	% of overall difference in gene number
GPCRs	1465	326	1139	16
F-box proteins	381	34	347	5
NHRs	276	87	189	3
C-type lectins	162	51	111	2
MSPs	108	50	58	1
Total	*2392*	*548*	*1844*	*26*

Table S3: Large gene families in *C. bovis* and *C. elegans*. Related to Figure 4.
Gene families that are known to make up a substantial fraction of the *C. elegans* genome [86] were chosen.

	C. bovis	*C. elegans*
Chromosome I (Mb / %)	1.5 / 16	3.0 / 20
Chromosome II (Mb / %)	1.4 / 14	2.5 / 16
Chromosome III (Mb / %)	1.5 / 16	2.7 / 19
Chromosome IV (Mb / %)	1.6 / 16	2.6 / 15
Chromosome V (Mb / %)	1.6 / 15	3.5 / 17
Chromosome X (Mb / %)	0.6 / 5	2.1 / 12
Total (Mb / %)	8.2 / 13	16.3 / 16

Table S4: Repetitive content in the *C. bovis* and *C. elegans* genomes. Related to Figure 4. Repeat content for *C. bovis* chromosomes was estimated by summing the estimated repeat content of the 15 contigs classified into chromosomes in Figure S4.

	All genes		Orthologues only	
	C. bovis	*C. elegans*	*C. bovis*	*C. elegans*
gene count	13128	20208	7706	7706
total span of genes (excluding UTRs) (bp)	35062444	57199834	21555520	30210405
total exon span (bp)	18715784	24639393	11766915	11906370
mean exon span per gene (bp)	1426	1219	1527	1545
total exon count	117212	122373	73985	58546
mean exon count per gene	8.93	6.06	9.60	7.60
total intron span (bp)	16346660	32560441	9788605	18304035
mean intron span per gene (bp)	1245	1611	1270	2375
total intron count	104084	102165	66279	50840
mean intron count per gene	7.93	5.06	8.60	6.60
mean intron length (bp)	157	319	148	360

Table S4: Repetitive content in the *C. bovis* and *C. elegans* genomes. Related to Figure 4. Counts shown are for the longest isoform of each gene. UTRs were not annotated for *C. bovis* and so were not considered in either species. WormBase ParaSite version WBPS12 of the *C. elegans* genome was used.

Appendix C: Supplementary material for chapter 4

Table S1: gCF, branch lengths and distances from root for all recovered branches

Node #	*gCF*	*Branch length*	*Distance from root*
1	99.8	0.3847175012	0.3847175012
2	99.8	0.3847175012	0.3847175012
3	44.2	0.0589471269	0.4436646281
4	71.2	0.1361229852	0.5797876133
5	58.7	0.0511507553	0.6309383686
6	32.5	0.0320565182	0.6118441315
7	6.33	0.0130912995	0.624935431
8	6.35	0.0084674743	0.6334029053
9	4.69	0.0076190653	0.6325544963
10	32.8	0.0254552767	0.658858182
11	91.4	0.0812507262	0.7146536315
12	95.6	0.1362871264	0.7688416227
13	45.3	0.0238151637	0.65636966
14	90.3	0.080888803	0.739746985
15	89.6	0.0425405806	0.7571942121
16	95.7	0.0716259028	0.8404675255
17	81.2	0.0347404684	0.8035820911
18	56.6	0.0349455185	0.6913151785
19	86.8	0.069370048	0.725739708
20	97.7	0.0951025528	0.8348495378
21	81.1	0.0239042316	0.8274863227
22	93.5	0.0440225589	0.84760465
23	97.4	0.0932627989	0.7845779774
24	64	0.0162393301	0.7419790381
25	20.8	0.0069846302	0.7327243382
26	32.2	0.0032513831	0.8307377058
27	81.6	0.0116598055	0.8592644555
28	48.9	0.0104388887	0.7524179268
29	95.7	0.0600937209	0.802072759
30	64.5	0.0238949012	0.7566192394
31	23.2	0.006446124	0.7391704622
32	25.7	0.0045279531	0.7569458799
33	31	0.0080411276	0.7604590544
34	84.7	0.0320161361	0.7886353755
35	73	0.024353834	0.7635242962
36	92.2	0.0550998264	0.7942702886
37	37.1	0.0075542782	0.7645001581
38	62.9	0.012571034	0.8012064095
39	53.9	0.0115690595	0.7750933557
40	48.9	0.0114359782	0.7759361363
41	38.6	0.0075129856	0.7720131437
42	70.7	0.0151164798	0.7902098355
43	48.3	0.0093290823	0.784422438
44	64.4	0.0154456696	0.7913818059
45	24.3	0.0043809934	0.7803171297
46	64	0.0154855653	0.787498709
47	36.5	0.0064996104	0.7785127541
48	75	0.0126551085	0.802864944
49	34.7	0.0059764916	0.7903989296
50	87	0.0231817112	0.8145635171
51	30.7	0.0039532585	0.7953350644
52	72.3	0.0178794703	0.7981966
53	93.6	0.0355761369	0.8230748459
54	70.8	0.0169841778	0.8044828868
55	84.6	0.0226799163	0.8011926704
56	36.3	0.0035799499	0.8080628367
57	80.7	0.0102327451	0.8114254155
58	62.4	0.0083755508	0.8164383875

Table S2: Software, versions and relevant parameters used in phylogenomic analysis

Software	Version	Relevant parameters
OrthoFinder	2.2.7	-og
NCBI-BLAST+	2.5.0	-evalue 1e-5 -outfmt '6' -seg yes -soft_masking true -use_sw_tback
KinFin	1.0	
MAFFT	v7.407	--auto
IQ-TREE	1.6.10	-bb 1000 -bb 1000 -m GTR20+G
PhyloTreePruner	V20150918	0.9
trimAl	v1.4.rev15	-gt 0.8 -st 0.001 -resoverlap 0.75 -seqoverlap 80
catfasta2phyml	v1	-f -c
PhyloBayes	1.8 (mpi)	-cat -dp -gtr -dgam 4
ASTRAL-III	5.6.3	
Tracer	v1.7.1	

Table S3: Details and accessions of all data used in phylogenomic analyses

Species	INSDC Accession	Source
C. afra	-	-
C. angaria	PRJNA51225	WormBase ParaSite
C. astrocarva	-	-
C. becei	PRJEB28243	Luke Noble
C. bovis	-	-
C. brenneri	PRJNA20035	WormBase ParaSite
C. briggsae	PRJNA10731	WormBase ParaSite
C. castelli	-	-
C. dolens	-	-
C. doughertyi	-	-
C. drosophilae	-	-
C. elegans	PRJNA13758	WormBase ParaSite
C. guadeloupensis	-	-
C. imperialis	-	-
C. inopinata	PRJDB5687	WormBase ParaSite
C. japonica	PRJNA12591	WormBase ParaSite
C. kamaaina	-	Luke Noble
C. latens	PRJNA248912	WormBase ParaSite
C. macrosperma	-	-
C. monodelphis	-	-
C. nigoni	PRJNA384657	WormBase ParaSite
C. nouraguensis	-	-
C. oiwi	-	-
C. panamensis	PRJEB28259	Luke Noble
C. parvicauda	PRJEB12595	-
C. plicata	-	-
C. portoensis	-	-
C. quiockensis	PRJEB11354	-
C. remanei	PRJNA248911	WormBase ParaSite
C. sinica	PRJNA194557	WormBase ParaSite
***C.* sp. 2**	-	-
***C.* sp. 24**	-	-
***C.* sp. 25**	-	-
***C.* sp. 27**	-	-
***C.* sp. 30**	-	-
***C.* sp. 33**	-	John Wang
***C.* sp. 41**	-	John Wang
***C.* sp. 44**	-	John Wang
***C.* sp. 45**	-	-
***C.* sp. 46**	-	-
***C.* sp. 47**	-	-
***C.* sp. 48**	-	-
***C.* sp. 49**	-	-
***C.* sp. 51**	-	-
***C.* sp. 54**	-	-
***C.* sp. 55**	-	-
***C.* sp. 56**	-	-
***C.* sp. 8**	-	-
C. sulstoni	PRJEB12601	-
C. tribulationis	PRJEB12608	-
C. tropicalis	PRJNA53597	WormBase ParaSite
C. uteleia	PRJEB12600	-
C. virilis	-	-
C. vivipara	-	-
C. waitukubuli	PRJEB12602	-
C. wallacei	-	-
C. yunquensis	-	Christian Riccio
C. zanzibari	PRJEB12596	-
Diploscapter coronatus	PRJDB3143	WormBase ParaSite
Diploscapter pachys	PRJNA280107	WormBase ParaSite

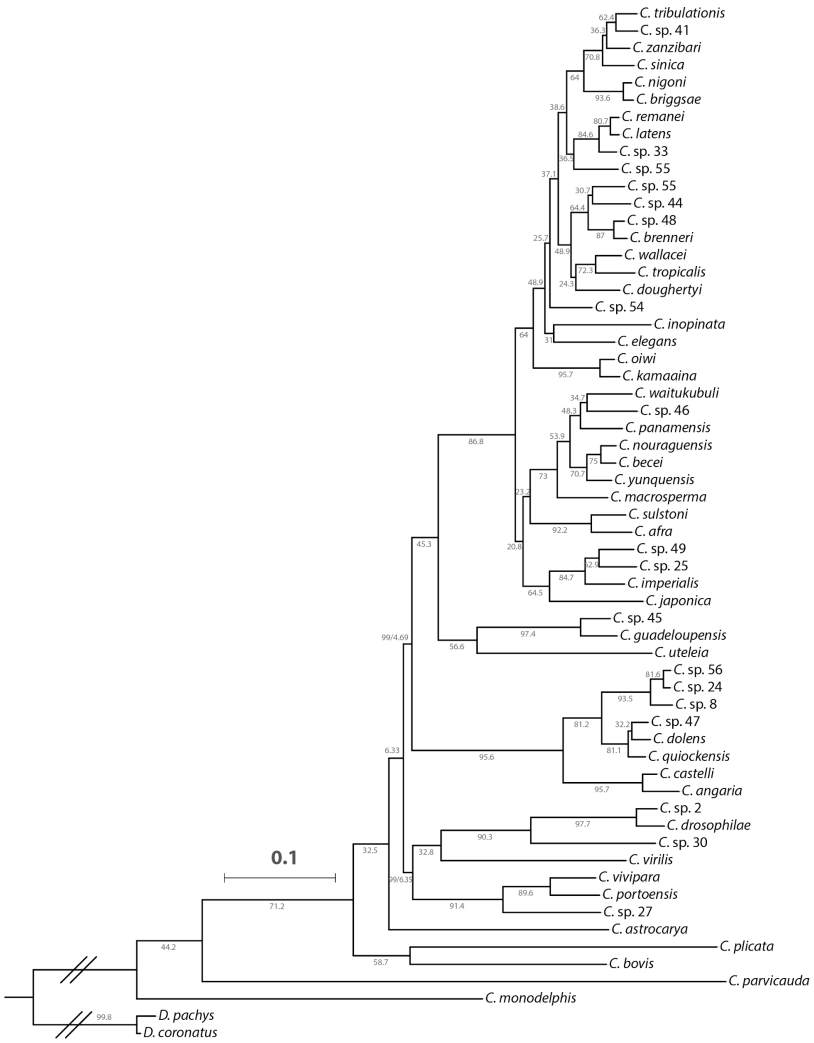

Figure S1: Phylogenetic tree of 58 *Caenorhabditis* species and two outgroup taxa using maximum likelihood

Phylogenetic tree inferred using IQ-TREE with the GTR+Γ substitution model. A supermatrix containing 2,869 single-copy orthologues was used. The proportion of gene trees concordant with each branch (gCF) are noted. UFBootstrap support values (1000 replicates) are 100 unless noted as branch annotations (BS/gCF). Scale is in substitutions per site.

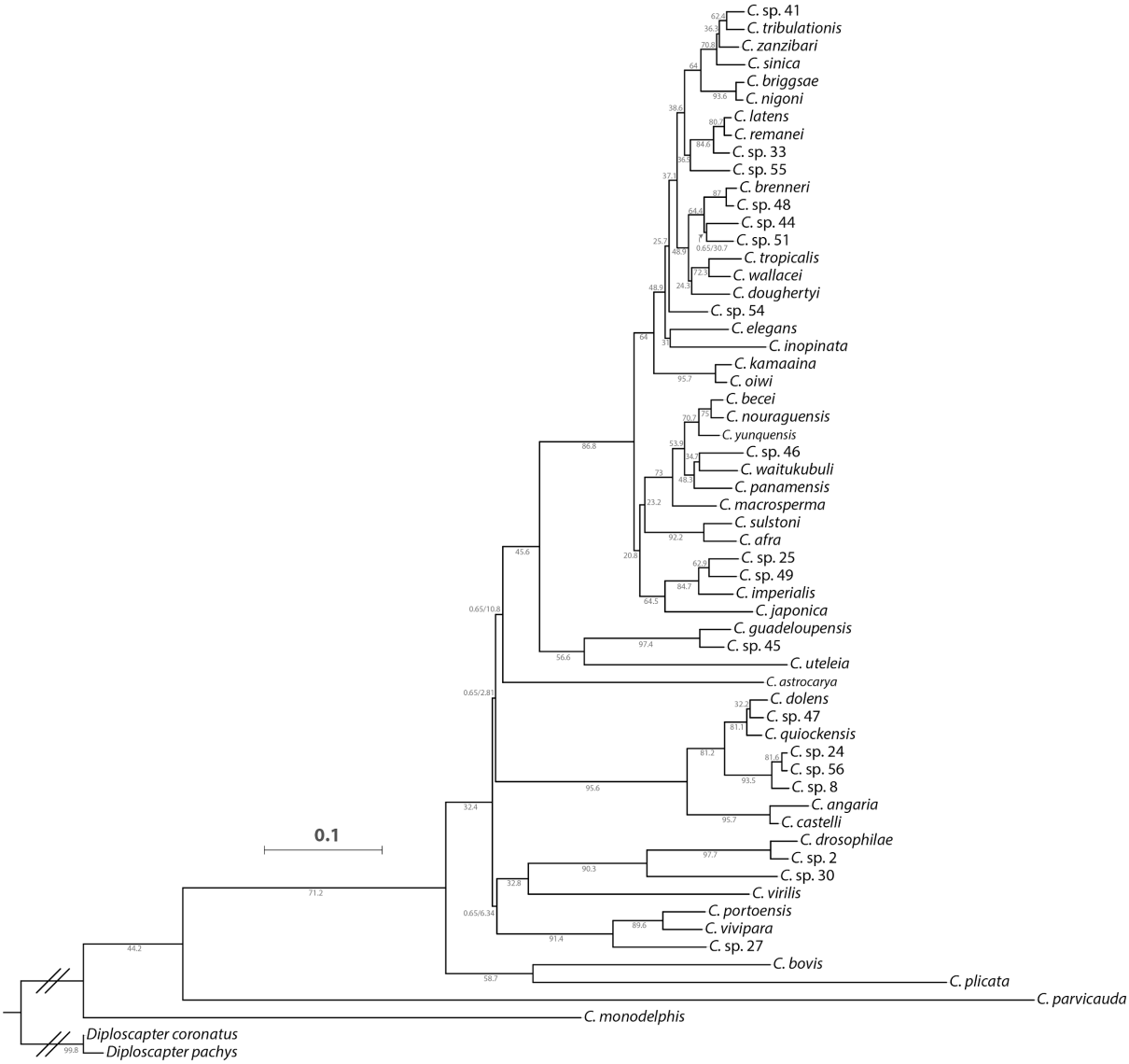

Figure S2: Phylogenetic tree of 58 *Caenorhabditis* species and two outgroup taxa inferred using Bayesian inference

Phylogenetic tree inferred using PhyloBayes with the CAT-GTR+Γ substitution model. A reduced supermatrix containing 467 orthologues was used. The proportion of gene trees concordant with each branch (gCF) are noted. Posterior probabilities are 1.0 unless noted as branch annotations (PP/gCF). Scale is in substitutions per site.

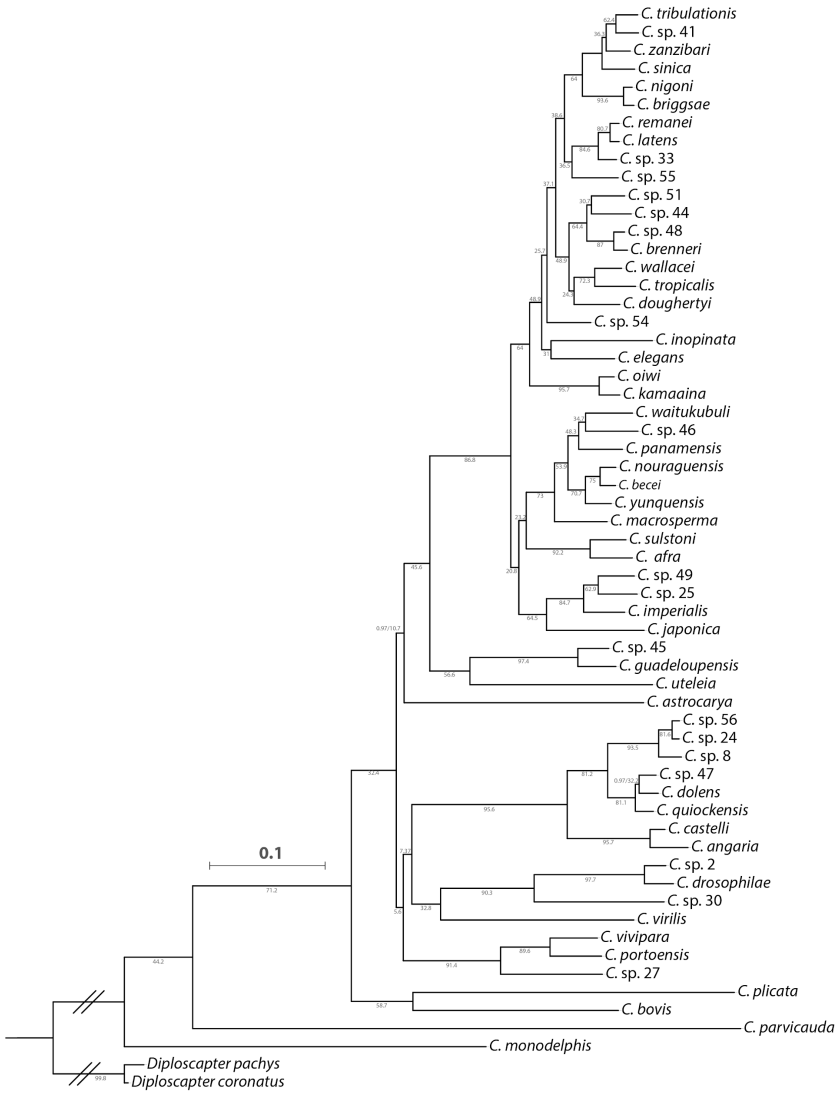

Figure S3: Phylogenetic tree of 58 *Caenorhabditis* species and two outgroup taxa inferred using ASTRAL-III

Phylogenetic tree inferred using ASTRAL-III, by providing maximum likelihood gene trees (inferred using IQ-TREE with the substitution model selected automatically) as input. As ASTRAL-III outputs trees with branch lengths in coalescent units, branch lengths in substitutions per site were estimated using IQ-TREE with the GTR+Γ substitution model and the concatenated alignment. The proportion of gene trees concordant with each branch (gCF) are noted. Posterior probabilities are 1.0 unless noted as branch annotations (PP/gCF). Scale is in substitutions per site.

Appendix D: Supplementary material for chapter

Table S1: Details and accessions of all data used in chapter 5

Species	Used in genome size analysis? (Y/N)	INSDC Accession	Source
C. afra	Y	-	-
C. angaria	N	PRINA51225	WormBase ParaSite
C. astrocarva	Y	-	-
C. becei	Y	PRIEB28243	Luke Noble
C. bovis	Y	-	-
C. brenneri	N	PRINA20035	WormBase ParaSite
C. briggsae	Y	PRINA10731	WormBase ParaSite
C. castelli	Y	-	-
C. dolens	Y	-	-
C. doughertyi	Y	-	-
C. drosophilae	Y	-	-
C. elegans	Y	PRINA13758	WormBase ParaSite
C. guadeloupensis	N	-	-
C. imperialis	Y	-	-
C. inopinata	Y	PRIDB5687	WormBase ParaSite
C. japonica	N	PRINA12591	WormBase ParaSite
C. kamaaina	Y	-	Luke Noble
C. latens	Y	PRINA248912	WormBase ParaSite
C. macrosperma	Y	-	-
C. monodelphis	Y	-	-
C. nigoni	Y	PRINA384657	WormBase ParaSite
C. nouraguensis	Y	-	-
C. oiwi	Y	-	-
C. panamensis	Y	PRIEB28259	Luke Noble
C. parvicauda	Y	PRIEB12595	-
C. plicata	Y	-	-
C. portoensis	Y	-	-
C. quiockensis	Y	PRIEB11354	-
C. remanei	Y	PRINA248911	WormBase ParaSite
C. sinica	Y	PRINA194557	WormBase ParaSite
C. sp. 2	Y	-	-
C. sp. 24	Y	-	-
C. sp. 25	Y	-	-
C. sp. 27	Y	-	-
C. sp. 30	Y	-	-
C. sp. 33	Y	-	John Wang
C. sp. 41	Y	-	John Wang
C. sp. 44	Y	-	John Wang
C. sp. 45	Y	-	-
C. sp. 46	N	-	-
C. sp. 47	Y	-	-
C. sp. 48	Y	-	-
C. sp. 49	N	-	-
C. sp. 51	Y	-	-
C. sp. 54	Y	-	-
C. sp. 55	Y	-	-
C. sp. 56	Y	-	-
C. sp. 8	Y	-	-
C. sulstoni	Y	PRIEB12601	-
C. tribulationis	Y	PRIEB12608	-
C. tropicalis	Y	PRINA53597	WormBase ParaSite
C. uteleia	Y	PRIEB12600	-
C. virilis	Y	-	-
C. vivipara	N	-	-
C. waitukubuli	N	PRIEB12602	-
C. wallacei	Y	-	-
C. vunaensis	Y	-	Christian Riccio
C. zanzibari	Y	PRIEB12596	-
Diploscapter coronatus	N	PRIDB3143	WormBase ParaSite
Diploscapter pachys	N	PRINA280107	WormBase ParaSite

Table S2: Software, versions and relevant parameters used in phylogenomic analysis

Software	Version	Relevant Parameters
OrthoFinder	v2.2.7	
KinFin	v1.0	
InterProScan	5.35-74.0	-dp -t p --goterms -appl SignalP_EUK,Pfam
RepeatModeller	open-1.0.11	open-4.0.9
RepeatMasker	open-4.0.9	-engine ncbi
ape (R Package)	5.2	
caper (R Package)	1.0.1	
phytools (R package)	0.6-60	
geiger (R package)	2.0.6	
nlme (R package)	3.1-137	
MAFFT	v7.407	--auto
IQ-TREE	1.6.10	-bb 1000 (substitution model automatically selected)
PhyloTreePruner	v20150918	0.9 u
PAL2NAL	v14	-fasta

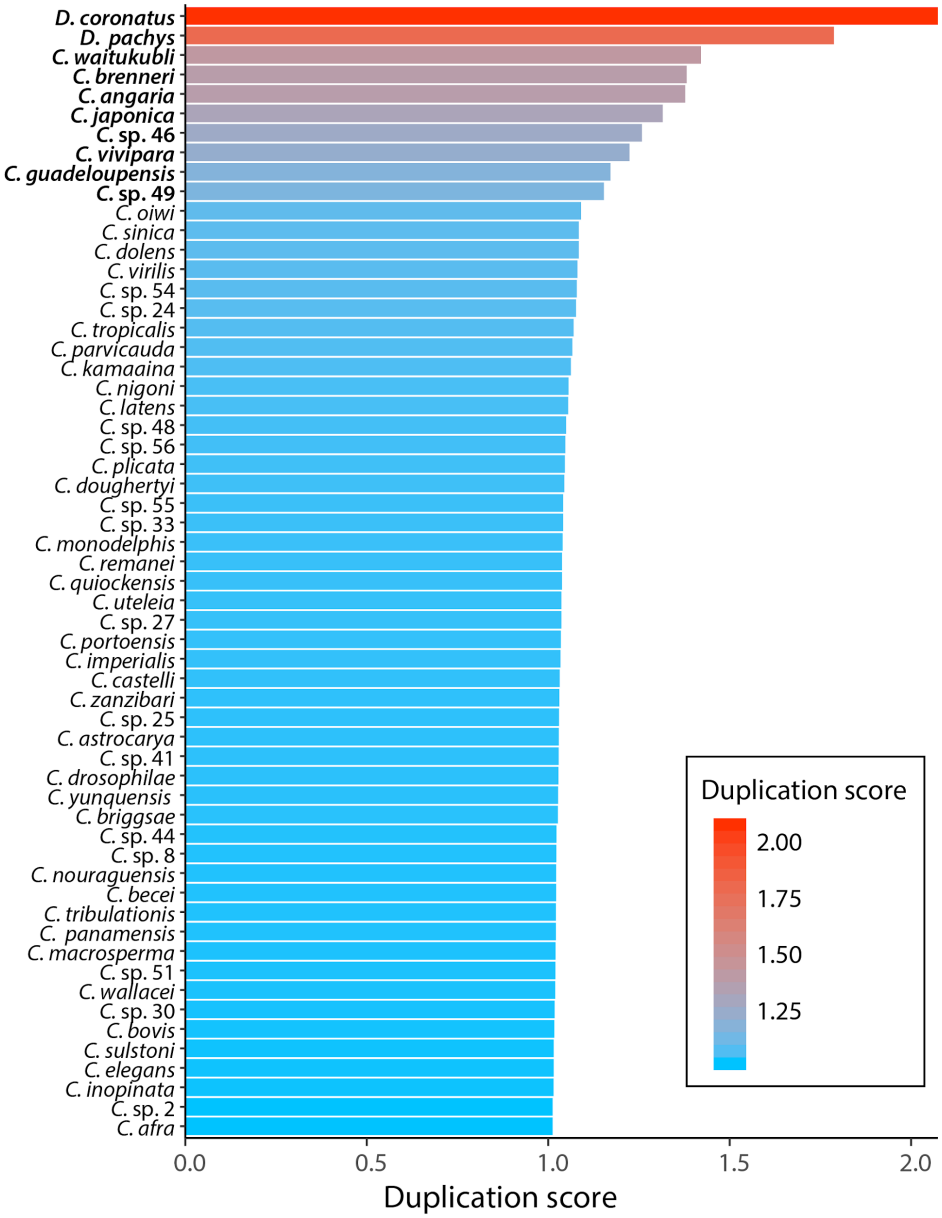

Figure S1: Duplication in the genomes of 58 *Caenorhabditis* species and two *Diploscapter* species

Duplication score was calculated by dividing the total number of genes present in 1,987 orthogroups that were, on average, single-copy in all 60 species. Bars are coloured by duplication score. Species highlighted in bold were excluded.

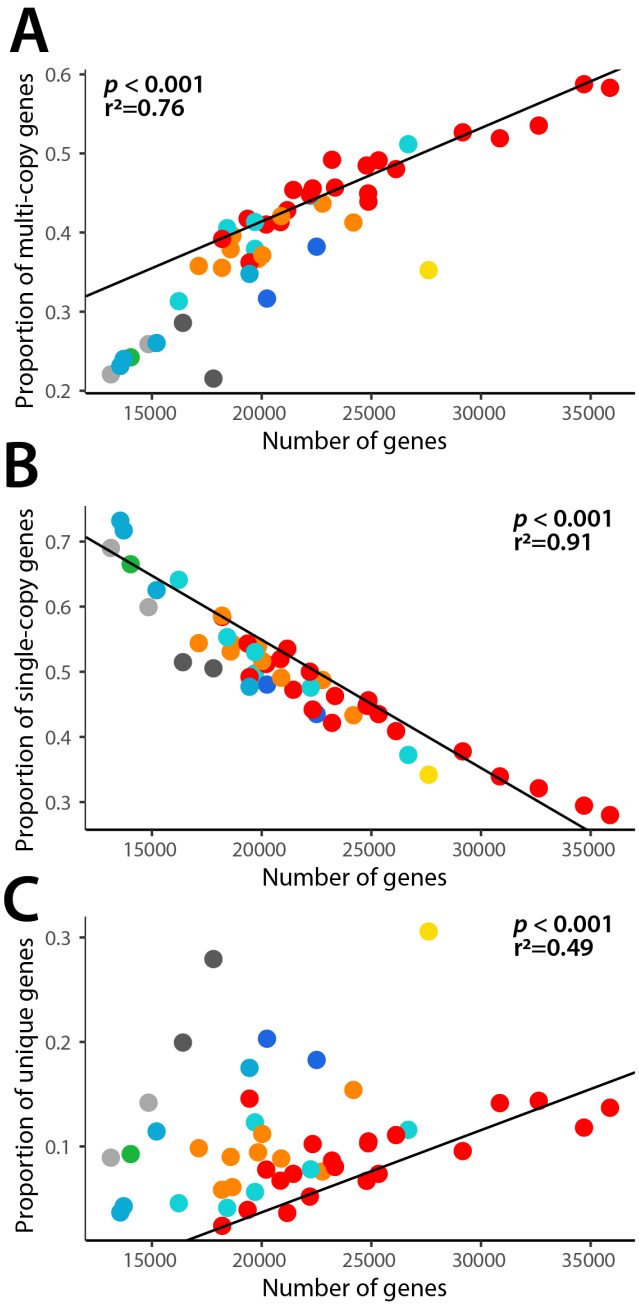

Figure S2: PGLS analysis of protein-coding gene number

A: Proportion of multi-copy genes versus protein-coding gene number. B: Proportion of single-copy genes protein-coding gene number. C: Proportion of unique genes protein-coding gene number. PGLS was performed using the ape, caper, and phytools R packages using the Brownian model of evolution and the phylogenetic tree in Figure 1A.

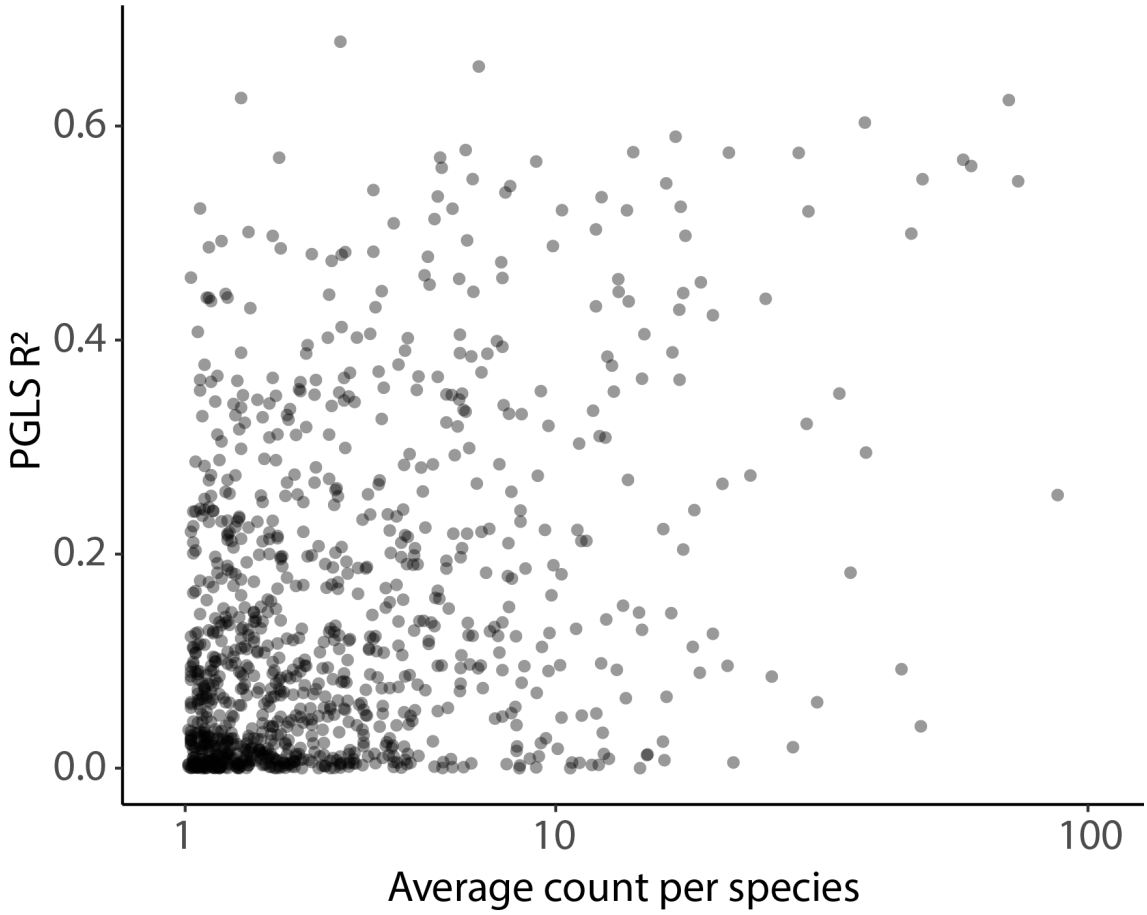

Figure S3: Gene family size correlated with protein-coding gene number

Counts for each species in each gene family were calculated using KinFin. These counts were compared with the total number of protein-coding genes in each species. R^2 values were calculated PGLS using the ape, caper, and phytools R packages with the Brownian model of evolution and the phylogenetic tree in Figure 1A.